

4

DTIC FILE COPY

AD-A196 071

CLaSSiC Project  
Manuscript CLaSSiC-88-22

May 1988

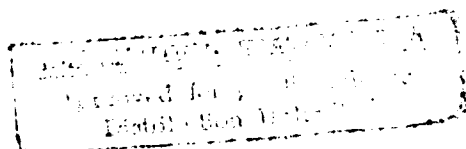
# Algorithms for Nonlinear Least-Squares Problems

CONTRACT N00014-82-K-0335

Christina Fraley



Center for Large Scale Scientific Computation  
Building 460, Room 313  
Stanford University  
Stanford, California 94305-2140



- 1 -

**Christina Fraley**

## Abstract

This paper addresses the nonlinear least-squares problem  $\min_{x \in \mathbb{R}^n} \|f(x)\|_2^2$ , where  $f(x)$  is a vector in  $\mathbb{R}^m$  whose components are smooth nonlinear functions. The problem arises most often in data fitting applications. Much research has focused on the development of specialized algorithms that attempt to exploit the structure of the nonlinear least-squares objective. We survey methods developed for problems in which sparsity in the derivatives of  $f$  is not taken into account in formulating algorithms.

1. Monday - 10th - Trade,  
 2. Tuesday - 11th - Trade, Good at  
evening, after the at, the



Account No.	J
NTIS Order #	
Delivery Date	12
User Name	12
Justification	
By <i>per ltr</i>	
Date	
Comments	
Author	
Subject	
A-1	

### Acknowledgements

I would like to thank David Gay, Philip Gill, and Margaret Wright for many helpful comments and suggestions. I would also like to acknowledge generous financial support from the following sources :

- the Xerox Corporation, in the form of a fellowship.
- Joseph Oliger, under Office of Naval Research contract # N00014-82-K-0335,
- Jean-Philippe Vial, under Fonds National de la Recherche Scientifique Suisse research grant # 1 467-0.86,
- the Systems Optimization Laboratory, Stanford University, under National Science Foundation grant # CCR-8413211.

## 1. Introduction

This paper addresses the problem of minimizing the  $l_2$  norm of a multivariate function :

$$\min_{x \in \mathbb{R}^n} \frac{1}{2} \|f(x)\|_2^2,$$

where  $f(x)$  is a vector in  $\mathbb{R}^m$  whose components are real-valued nonlinear functions with continuous second partial derivatives. We shall refer to the function  $\frac{1}{2} \|f(x)\|_2^2$  as the nonlinear *least-squares objective function*. An alternative formulation of the problem is that of minimizing a sum of squares :

$$\min_{x \in \mathbb{R}^n} \frac{1}{2} \sum_{i=1}^m \phi_i(x)^2,$$

where each  $\phi_i$  is a real-valued function having continuous second partial derivatives.

There is considerable interest in the nonlinear least-squares problem, because it arises in virtually all areas of quantitative research in data-fitting applications. A typical instance is the choice of parameters  $\beta$  within a nonlinear model  $\varphi$  so that the model agrees with measured quantities  $d_i$  as closely as possible :

$$\min_{\beta \in \mathbb{R}^n} \sum_{i=1}^m \frac{1}{2} (\varphi(\beta; \tau_i) - d_i)^2,$$

where  $\tau_i$  are prescribed values. Much research has focused on the development of specialized algorithms that attempt to exploit the structure of the nonlinear least-squares objective. Despite these efforts, methods do not perform equally well on all problems, and it is generally not possible to characterize those problems on which a particular method will work well.

In this paper, we survey existing numerical methods for dense nonlinear least-squares problems. For a study of the performance of widely-distributed software for nonlinear least-squares, see Fraley [1987a, 1988]. We assume a knowledge of numerical methods for linear least-squares problems (e. g., Lawson and Hanson [1974], and Golub and Van Loan [1983]). We also assume familiarity with Newton-based linesearch and trust-region methods for unconstrained minimization (e. g., Fletcher [1980], Gill, Murray, and Wright [1981], Dennis and Schnabel [1983], and Moré and Sorensen [1984]). If  $\mathcal{F}$  is the function to be minimized, recall that both linesearch and trust-region methods involve iterative minimization of a quadratic local model

$$Q(p) = \nabla \mathcal{F}(x_k)^T p + \frac{1}{2} p^T H_k p$$

for  $\mathcal{F}(x_k + p) - \mathcal{F}(x_k)$ , the change in  $\mathcal{F}$  at the current iterate  $x_k$ . In linesearch methods, the vector  $p_k^{LS}$  defined by

$$p_k^{LS} \equiv \arg \min_{p \in \mathbb{R}^n} Q(p)$$

is used as a *search direction*. A positive step is taken from  $x_k$  along  $p_k^{LS}$  to the next iterate, that is,

$$x_{k+1} = x_k + \alpha_k p_k^{LS},$$

where the *steplength*  $\alpha_k > 0$  is computed by approximate minimization of the function  $\Phi_k(\alpha) = \mathcal{F}(x_k + \alpha p_k^{LS})$ . The vector  $p_k^{LS}$  must be a *descent direction* for  $\mathcal{F}$  at  $x_k$  — in other words,  $\nabla \mathcal{F}(x_k)^T p_k^{LS} < 0$  — so that  $\mathcal{F}$  initially decreases along  $p_k^{LS}$  from  $x_k$ . Normally  $H_k$  is required to be positive definite, which guarantees that the quadratic model has a unique minimum that is a descent direction. In trust-region methods,

$$x_{k+1} = x_k + p_k^{TR},$$

where

$$p_k^{TR} = \arg \min_{p \in \mathbb{R}^n} Q(p) \quad \text{subject to} \quad \|p\| \leq \delta_k.$$

The rationale for restricting the size of  $p$  in the subproblem is that  $Q(p)$  is a good approximation to  $\mathcal{F}$  only at points close to  $x_k$ .

## 1.1 Definitions and Notation

We shall use the following definitions and notational conventions:

- Generally subscripts on a function mean that the function is evaluated at the corresponding subscripted variable (for example,  $f_k = f(x_k)$ ). An exception is made for the residual functions  $\phi_i$ , where the subscript is the component index for the vector  $f$ .
- $f$  - The vector of nonlinear functions whose  $l_2$  norm is to be minimized.  
The nonlinear least-squares problem is

$$\min_{x \in \mathbb{R}^n} \frac{1}{2} f(x)^T f(x),$$

where the factor  $\frac{1}{2}$  is introduced in order to avoid a factor of two in the derivatives.

- $\phi_i$  - The  $i$ th *residual* function, also the  $i$ th component of the vector  $f$ .

$$f(x) \equiv \begin{pmatrix} \phi_1(x) \\ \vdots \\ \phi_m(x) \end{pmatrix}.$$

An alternative formulation of the nonlinear least-squares problem is

$$\min_{x \in \mathbb{R}^n} \frac{1}{2} \sum_{i=1}^m \phi_i(x)^2.$$

where each  $\phi_i(x)$  is a smooth function mapping  $\mathbb{R}^n$  to  $\mathbb{R}$ .

- $J$  - The  $m \times n$  Jacobian matrix of  $f$ .

$$J(x) \equiv \nabla f(x) = \begin{pmatrix} \frac{\partial \phi_1}{\partial x_1} & \cdots & \frac{\partial \phi_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial \phi_m}{\partial x_1} & \cdots & \frac{\partial \phi_m}{\partial x_n} \end{pmatrix}$$

- $g$  - The gradient of the nonlinear least-squares objective.

$$g(x) \equiv \nabla \left( \frac{1}{2} f(x)^T f(x) \right) = J(x)^T f(x)$$

- $B$  - The part of the Hessian matrix of the nonlinear least-squares objective that involves second derivatives of the residual functions. We have

$$\nabla^2 \left( \frac{1}{2} f(x)^T f(x) \right) = J(x)^T J(x) + B(x),$$

where

$$B(x) \equiv \sum_{i=1}^m \phi_i(x) \nabla^2 \phi_i(x).$$

- $\mathcal{R}(A)$  - The range of  $A$ .

If  $A$  is an  $m \times n$  matrix, then  $\mathcal{R}(A) \equiv \{b \in \mathbb{R}^m \mid Ax = b \text{ for some } x \in \mathbb{R}^n\}$  is a subspace of  $\mathbb{R}^m$ .

- $\mathcal{N}(A)$  - The null space of  $A$ .

If  $A$  is an  $m \times n$  matrix, then  $\mathcal{N}(A) \equiv \{z \in \mathbb{R}^n \mid Az = 0\}$  is a subspace of  $\mathbb{R}^n$ .  $\mathcal{N}(A)$  is the orthogonal complement of  $\mathcal{R}(A^T)$  in  $\mathbb{R}^n$ .

## 2. Gauss-Newton Methods

The classical approach to nonlinear least squares, called the Gauss-Newton method, is a linesearch method in which the search direction at the current iterate minimizes the quadratic function

$$g_k^T p + \frac{1}{2} p^T J_k^T J_k p. \quad (2.1)$$

The function (2.1) is a local approximation to  $\frac{1}{2} \|f(x_k + p)\|_2^2 - \frac{1}{2} \|f(x_k)\|_2^2$  in which each residual component of  $f$  is approximated by a linear function, using the relationship

$$f(x_k + p) = f(x_k) + J(x_k)p + \mathcal{O}(\|p\|^2).$$

As a model for the change in the least-squares objective, (2.1) has the advantage that it involves only first derivatives of the residuals, and that  $J^T J$  is always at least positive semi-definite. If

$$p_k^{GN} \equiv \arg \min_{p \in \mathbb{R}^n} g_k^T p + \frac{1}{2} p^T J_k^T J_k p,$$

then  $p_k^{GN}$  satisfies the equations

$$J_k^T J_k p = -g_k. \quad (2.2)$$

and is therefore a direction of descent for  $f_k^T f_k$  whenever  $g_k \neq 0$  — as required in a linesearch method. To guarantee convergence to a local minimum, the sequence of search directions must also be bounded away from orthogonality to the gradient, a condition that may not be met by successive Gauss-Newton directions unless the eigenvalues of  $J^T J$  are bounded away from zero. Powell [1970] gives an example of convergence of a Gauss-Newton method with exact linesearch to a non-stationary point.

The Gauss-Newton method can be viewed as a modification of Newton's method in which  $J^T J$  is used to approximate the Hessian matrix

$$J^T J + \sum_{i=1}^m \phi_i \nabla^2 \phi_i = J^T J + B$$

of the nonlinear least-squares objective function. The assumption is that the matrix  $J^T J$  should be a good approximation to the full Hessian when the residuals are small. In fact, if  $f(x^*) = 0$  and  $J(x^*)^T J(x^*)$  is positive definite, then the sequence  $\{x_k + p_k^{GN}\}$  is locally quadratically convergent to  $x^*$ , because

$$J(x_k)^T J(x_k) = \frac{1}{2} \nabla^2 \|f(x^*)\|_2^2 + \mathcal{O}(\|x_k - x^*\|).$$

For more convergence results and detailed convergence analysis for the Gauss-Newton method, see, e. g., Chapter 10 of Dennis and Schnabel [1983], Schaback [1985], and Häussler [1986], as well as some of the references cited below.

McKeown [1975a, 1975b] studies test problems of the form,

$$f(x) = f_0 + G_0 x + \frac{1}{2} \begin{pmatrix} x^T H_1 x \\ \vdots \\ x^T H_m x \end{pmatrix},$$

chosen so that factors affecting the rate of convergence could be controlled. He uses three such problems, with seven different values of a parameter that varies an asymptotic linear convergence factor. The algorithms tested include some quasi-Newton methods for unconstrained optimization, as well as some specialized methods for nonlinear least squares that have since

been superseded. He concludes that, when the asymptotic convergent factor is small, the Gauss-Newton method is more efficient than the quasi-Newton methods, but that the opposite is true when the asymptotic convergence factor is large. Fraley [1987a, b; 1988] gives numerical results for some Gauss-Newton methods using these problems, and observes that the Jacobian is well-conditioned at every iteration.

A difficulty with the Gauss-Newton method arises when  $J^T J$  is singular, or, equivalently, when  $J$  has linearly dependent columns, because then (2.1) does not have a unique minimizer. For this reason the Gauss-Newton method should more accurately be viewed as a class of methods, each member being distinguished by a different choice of  $p$  when  $J^T J$  is singular. The set of vectors that minimize (2.1) is the same as the set of solutions to the linear least-squares problem

$$\min_{p \in \mathbb{R}^n} \|J_k p + f_k\|_2. \quad (2.3)$$

One (theoretically) well-defined alternative that is often approximated computationally is to require the unique solution of minimum  $l_2$  norm:

$$\min_{p \in S} \|p\|_2, \quad (2.4)$$

where  $S$  is the set of solutions to (2.3). Another alternative is to replace  $J$  in (2.3) by a maximal linearly independent subset of its columns. In finite-precision arithmetic, there is often some ambiguity about how to formulate and solve an alternative to (2.3) when the columns of  $J$  are "nearly" linearly dependent, so that, from a computational standpoint, any particular Gauss-Newton method must be still viewed as a class of methods. The references cited above for linear least squares discuss at length the difficulties inherent in computing solutions to (2.3) when  $J$  is ill-conditioned, and show that the numerical solution of these problems is dependent on the criteria used to estimate the rank of  $J$ . From now on, the term "Gauss-Newton method" will refer to any linesearch method that has the following two properties. First, the nonlinear least-squares objective is used as a merit function for the linesearch. Second, the search direction is the result of some well-defined computational procedure for solving (2.3).

For a survey of some of the early research on numerical Gauss-Newton methods, see Dennis [1977]. More recently, Deuffhard and Apostolescu [1980] suggest selecting a steplength for the Gauss-Newton direction based on decreasing the merit function  $\|J_k^\dagger f(x)\|_2^2$  rather than  $\|f(x)\|_2^2$ , for a class of nonlinear least-squares problems that includes zero-residual problems. The function  $J_k^\dagger$  is the pseudo-inverse of  $J_k$  (see, e. g., Chapter 6 of Golub and Van Loan [1983]);  $J_k^\dagger f_k$  is another way of representing the minimum  $l_2$ -norm solution to  $\|J_k p + f_k\|_2$ . They reason that the Gauss-Newton direction is the steepest-descent direction for the function  $\|J_k^\dagger f(x)\|_2^2$ , so that the geometry of the level surfaces defined by  $\|J_k^\dagger f(x)\|_2^2$  is more favorable to avoiding small steps



in the linesearch. A shortcoming of this approach (pointed out by the authors) is that there are no global convergence results. The merit function depends on  $x_k$ , so that a different function is being reduced at each step. Another difficulty is that, although the authors state that numerical experience supports selection of a steplength based on  $\|J_k^T f(x)\|_2^2$  for ill-conditioned problems, the transformation  $J_k^T$  is not numerically well-defined under these circumstances. Therefore neither the Gauss-Newton search direction, nor the merit function, is numerically well-defined when the columns of  $J_k$  are nearly linearly dependent.

There is another reason why it is difficult to say precisely what is meant by a "Gauss-Newton method" for a particular nonlinear least-squares problem. To see this, let  $Q(x)$  be an  $l \times m$  orthogonal matrix function on  $\mathbb{R}^n$ , that is,  $Q(x)^T Q(x) = I$  for all  $x$ . Then  $\|Q(x)f(x)\|_2^2 = \|f(x)\|_2^2$  for all  $x$ , and consequently the function  $\tilde{f} \equiv Qf$  defines the same nonlinear least-squares problem as  $f$ . The Jacobian matrix of  $\tilde{f}$  is  $\tilde{J} \equiv QJ + (\nabla Q)f$ , so that a minimizer of  $\|\tilde{J}p + \tilde{f}\|_2$  will ordinarily be different from a minimizer of  $\|Jp + f\|_2$ , unless  $Q(x)$  happens to be a constant transformation. However, if both  $Q$  and  $f$  have  $k$  continuous derivatives, then  $\nabla^i \|Q(x)f(x)\|_2^2 = \nabla^i \|f(x)\|_2^2$  for  $i = 1, 2, \dots, k$ . Letting  $W \equiv (\nabla Q)f$ , so that  $\tilde{J} = QJ + W$ , we have

$$\tilde{J}^T \tilde{J} = J^T J + (J^T Q^T W + W^T Q J) + W^T W,$$

showing that the Gauss-Newton approximation  $J^T J$  to the full Hessian matrix is changed when  $f$  is transformed by an orthogonal function that varies with  $x$ . Thus, with exact arithmetic, there are many Gauss-Newton methods corresponding to a given vector function, although Newton's method remains invariant (see also Nocedal and Overton [1985], p. 826). In fact, each step of a Gauss-Newton method could be defined by a different transformation of  $f$ . Moreover, the conditioning of  $\tilde{J}$  may be very different from that of  $J$ , so that, for example, the columns of  $\tilde{J}$  might be strongly independent, while  $J$  is nearly rank deficient. Since the number of rows in  $Q$  may be greater than  $n$ , it is possible to imbed the given nonlinear least-squares problem in a larger one. This suggests the (so far unexplored) idea of preconditioning a Gauss-Newton method at each step with an orthogonal function.

Although it is known that Gauss-Newton methods do not work well under all circumstances, it is not possible to say anything more precise about the method when considering large and varied sets of test problems. Gauss-Newton methods are of practical interest because there are many instances in which they work very well in comparison to other methods. In fact, most successful specialized approaches to nonlinear least-squares problems are based to some extent on Gauss-Newton methods and attempt to exploit this behavior whenever possible. However, it

is not hard to find cases where Gauss-Newton methods perform poorly, so that they cannot be successfully applied to general nonlinear least-squares problems without modification.

Fraley [1987a, 1988] gives numerical results for a large set of test problems using widely-distributed software for unconstrained optimization and nonlinear least squares. She also includes some Gauss-Newton methods that use LSSOL [Gill et al. (1986a)] to solve the linear least-squares subproblem (2.3). Her findings confirm that Gauss-Newton methods are often among the best available techniques for nonlinear least squares — especially for zero-residual problems — but that there are many cases in which they fail or are inefficient. Detailed examples are presented that illustrate some of the difficulties involved in characterizing those problems on which Gauss-Newton methods will or will not work well (see also Fraley [1987b]).

Many attempts have been made to define algorithms that depart from the Gauss-Newton strategy only when necessary. Bard [1970] compares some Gauss-Newton-based methods with a Levenberg-Marquardt method (Section 3) and some quasi-Newton methods for unconstrained optimization on a set of ten test problems from nonlinear parameter estimation. He uses the eigenvalue decomposition of  $J^T J$  to solve the normal equations (2.2). In order to ensure a positive-definite system, he modifies the eigenvalues if their magnitude falls below a certain threshold. In addition, his implementations include bounds on the variables that are enforced by adding a penalty term to the objective function. He finds that the Gauss-Newton-based methods are more efficient in terms of function and derivative evaluations than the quasi-Newton methods, but that there is no significant difference in the relative performance of the Gauss-Newton-based methods and the Levenberg-Marquardt method.

Betts [1976] proposes an algorithm that combines a Gauss-Newton method with a method in which the Gauss-Newton approximate Hessian  $J^T J$  is augmented by a quasi-Newton approximation to the second-order term  $B = \sum_{i=1}^m \phi_i \nabla^2 \phi_i$  in the nonlinear least-squares Hessian (see Section 5). The algorithm starts with a Gauss-Newton method, and then switches to the augmented Hessian when it is believed that the iterates are near the solution. The criterion for the switch is

$$\|p_k\|_2 < \epsilon (1 + \|x_k\|_2), \quad (2.5)$$

for some  $\epsilon < 1$ . Results are presented for the hybrid methods, as well as for the underlying Gauss-Newton method and special quasi-Newton method (see Section 5), on a set of eleven test problems. Betts concludes that the hybrid method is superior, especially on problems with nonzero residuals, although the results he lists in his tables do not all have the same value of  $\epsilon$  in (2.5). Another issue that is not clarified is the treatment of near-singularity or indefiniteness in the quadratic model in any of the methods tested. Also the test (2.5) may not necessarily

imply that the Gauss-Newton iterates are in the vicinity of a solution, and could instead indicate inefficiency in the Gauss-Newton method at some arbitrary point.

Ramsin and Wedin [1977] compare the performance of a Gauss-Newton-based method with that of a Levenberg-Marquardt method for nonlinear least squares and a quasi-Newton method for unconstrained optimization, both from the Harwell Library. The quasi-Newton routine required an initial estimate  $H_0$  of the Hessian matrix, and the choice  $H_0 = J(x_0)^T J(x_0)$  was made on the basis of preliminary tests that showed equal or better performance compared to  $H_0 = I$ . The test problems were constructed so that asymptotic properties could be monitored and are similar to those of McKeown [1975a, 1975b] mentioned above. In all cases considered, the Jacobian matrix had full column rank at the solution. The algorithm of Ramsin and Wedin uses the steepest-descent direction, rather than the Gauss-Newton direction, whenever the decrease in the objective is considered unacceptably small. The experiments involved variation of a large number of parameters. Ramsin and Wedin conclude that their Gauss-Newton-based method and the Levenberg-Marquardt method are identical when the asymptotic convergence factor is small, but that neither method is consistently better for large asymptotic convergence factors. Also, they find that in instances when the asymptotic convergence factor is large, the quasi-Newton method may be more efficient, although superlinear convergence of the quasi-Newton method was never observed. Ramsin and Wedin maintain that Gauss-Newton should not be used when (i) the current iterate  $x_k$  is close to the solution  $x^*$ , and the relative decrease in the size of the gradient is small, (ii)  $x_k$  is not near  $x^*$ , and the decrease in the sum of squares relative to the size of the gradient is small, or (iii)  $J_k$  is nearly rank-deficient. Conditions (i) and (ii) are indicators of inefficiency for any minimization algorithm; in general the problem of ascertaining the closeness of an iterate to a minimum is as difficult as solving the original problem. As for condition (iii), rapidly convergent Gauss-Newton methods may exist even if nearly rank-deficient Jacobians are encountered, but that it appears difficult to formulate a single rule for estimating the rank of the Jacobian that is satisfactory for all such problems (see Fraley [1987a, b]).

Wedin and Lindström [1987] have developed a hybrid algorithm for nonlinear least-squares that combines a Gauss-Newton method with a finite-difference Newton method. A Gauss-Newton search direction is computed at every iteration using a  $QR$  factorization. The rank estimation criteria are complicated, and search directions for several different estimates of the rank of  $J$  may be tried before a step is actually taken. A finite difference Newton step may be used when the steps along Gauss-Newton directions become small, and the iterates are judged to be close to a solution. In the algorithm, the decision about whether the iterates are close to the solution is based on the relation  $\|f_k\|_2 > \gamma \beta_k$  for some  $\gamma > 1$ , where  $\beta_k$  is the norm of the projection of  $f_k$  onto the range of  $J_k$ . The ratio  $\beta_k/\beta_{k-1}$  is used as an estimate of an asymptotic linear

convergence factor. They give numerical results for a set of thirty large-residual test problems constructed by Al-Baali and Fletcher [1985], and compare their results with those given by Al-Baali and Fletcher for two hybrid Gauss-Newton/BFGS methods (Section 9) and a version of NL2SOL (Section 5). Wedin and Lindström find that their method gives better overall results than the other methods, although their method does fail in three cases due to a finite-difference Hessian that is not positive definite.

### 3. Levenberg-Marquardt Methods

In Levenberg-Marquardt methods, the Gauss-Newton quadratic model (2.1) is minimized subject to a trust-region constraint. The step  $p$  between successive iterates solves

$$\min_{p \in \mathbb{R}^n} g^T p + \frac{1}{2} p^T J^T J p \quad (3.1)$$

$$\text{subject to } \|Dp\|_2 \leq \delta,$$

for some  $\delta > 0$  and some diagonal scaling matrix  $D$  with positive diagonal entries. Equivalently,  $p$  minimizes the quadratic model

$$g^T p + \frac{1}{2} p^T (J^T J + \lambda D^T D) p, \quad (3.2)$$

for some  $\lambda \geq 0$ . Since the matrix  $J^T J + \lambda D^T D$  is positive semidefinite, minimizers  $p_\lambda$  of (3.2) satisfy the equations

$$(J^T J + \lambda D^T D) p = -g = -J^T f, \quad (3.3)$$

which are the normal equations for the linear least-squares problem

$$\min_{p \in \mathbb{R}^n} \left\| \begin{pmatrix} J \\ \sqrt{\lambda} D \end{pmatrix} p - \begin{pmatrix} f \\ 0 \end{pmatrix} \right\|_2^2. \quad (3.4)$$

Hence a regularization method (e. g., Chapter 25 of Lawson and Hanson [1974], Eldén [1977, 1984], Varah [1979], and Gander [1981]) is being used to solve the linear least-squares problem (2.3) for the step to the next iterate.

The paper by Levenberg [1944] is the earliest known reference to methods of this type. Based on the observation that the unit Gauss-Newton step  $p_{GN}$  often fails to reduce the sum of squares when  $\|p_{GN}\|$  is not especially small, he suggests limiting the size of the search direction by solving a "damped" least-squares subproblem,

$$\min_{p \in \mathbb{R}^n} \omega(g^T p + \frac{1}{2} p^T J^T J p) + \|Dp\|_2^2, \quad (3.5)$$

in which a weighted sum of squares of linearized residuals and components of the search direction is minimized. He proves the existence of a value of  $\omega$  for which

$$\|f(x + p_\omega)\|_2 < \|f(x)\|_2,$$

where  $p_\omega$  solves (3.5), thus ensuring a reduction in the sum of squares for a suitable value of  $\omega$ . A major drawback is that no automatic procedure is given for obtaining  $\omega$ . Levenberg suggests computing the value of  $\|f(x + p_\omega)\|_2$  for several trial values of  $\omega$ , locating an approximate minimum graphically, and then repeating this procedure with the improved estimates until a satisfactory value of  $\omega$  is obtained, but precise criteria for accepting a trial value are not given. Two alternatives are proposed for the diagonal scaling matrix  $D$  in (3.5):  $D = I$ , because it minimizes the directional derivative  $g^T p_\omega$  for  $\omega = 0$ , and the square root of the diagonal of  $J^T J$ , based on empirical observations. The claim is that the new method solves a wider class of problems than methods that existed at that time, and that it does so with relative efficiency.

Somewhat later, a similar method was (apparently independently) proposed. Morrison [1960] considers a quadratic model

$$g^T p + \frac{1}{2} p^T H p, \quad (3.6)$$

in which either  $H = J^T J$  or  $H = \nabla^2 (f^T f)$  (in the latter case, it is implicitly assumed that  $\nabla^2 (f^T f)$  is positive semidefinite). He advocates minimizing (3.6) over a neighborhood of the current point as does Levenberg, because (3.6) may not be a good approximation to  $\frac{1}{2} (\|f(x + p)\|_2^2 - \|f(x)\|_2^2)$  if the minimizer  $p^*$  is large in magnitude, and consequently the sum of squares may not be reduced at  $x + p^*$ . (In Hartley [1961], a linesearch is used with the Gauss-Newton direction for the same reason.) Morrison proves that the solution  $p_\lambda$  to

$$\min_{p \in \mathbb{R}^n} g^T p + \frac{1}{2} p^T (H + \lambda D) p$$

for  $\lambda > 0$  is the constrained minimum of (3.6) on the sphere of radius  $\|D p_\lambda\|_2$ , and that  $\|p_\lambda\|_2 \rightarrow 0$  as  $\lambda \rightarrow \infty$ . In Morrison's method, the step bound  $\delta$  is the independent parameter, rather than  $\lambda$ . No specifications are given for either  $\delta$  or  $D$ , although it is implied that they can be chosen heuristically for a given problem. Instead of minimizing (3.6) subject to  $\|D p\|_2 \leq \delta$ , constraints of the form  $|d_i x_i| < \delta$  are imposed, and the resulting subproblem is then solved using the eigenvalue decomposition of  $H$ . Although the theory and methods apply for any positive semi-definite  $H$  in (3.6), no generalization to unconstrained minimization is mentioned.

Marquardt [1963] extended Morrison's work, showing that the vector  $p_\lambda$  that solves (3.3) becomes parallel to the steepest-descent direction as  $\lambda \rightarrow \infty$ , so that  $p_\lambda$  interpolates between the Gauss-Newton search direction,  $p_0$ , and the steepest-descent direction,  $p_\infty$ . He points out

that the method determines both the direction from the current iterate to the next one, and the distance between the iterates along that direction, and that increasing  $\lambda$  decreases the step length, while shifting the direction away from orthogonality to the gradient of the sum of squares. Marquardt's strategy controls  $\lambda$  automatically by multiplying or dividing the current value by a constant factor  $\nu$  greater than 1. He maintains that the minimum of the Gauss-Newton model should be taken over the largest possible neighborhood, that is, that  $\lambda$  should be chosen as small as possible, so as to achieve faster convergence by biasing the search direction toward the Gauss-Newton direction when Gauss-Newton methods would work well. Thus, at the  $k$ th iteration,  $\lambda_k = \lambda_{k-1}/\nu$  is tried first, and then increased if necessary by multiples of  $\nu$  until a reduction in the sum of squares is obtained. A shortcoming of this scheme is that  $\lambda$  is always positive, so that the constraint in (3.1) is active in every subproblem, and consequently a full Gauss-Newton step can never be taken. Also, no efficient method is given for solving (3.3) for different values of  $\lambda$ . Motivated by statistical considerations, Marquardt uses the diagonal of  $J^T J$  for the scaling matrix  $D$  (one of the alternatives proposed by Levenberg), and mentions that this scaling has been widely used as a technique for computing solutions to ill-conditioned linear least-squares problems.

Since the appearance of Marquardt's paper, and also that of Goldfeld, Quandt, and Trotter [1966], which independently proposed trust-region methods for general unconstrained optimization, much research has been directed toward improvements within the framework presented there. Bard [1970] takes the eigenvalue decomposition of  $J^T J$  at each iteration, so that (3.3) can be easily solved for several values of  $\lambda$ , and so that it will be known whether or not  $J^T J$  is singular. Bartels, Golub, and Saunders [1970] show how to use the SVD of  $J$  instead of the eigenvalue decomposition for the same purpose. They also give an algorithm for computing  $\lambda$  given  $\delta$  that involves determining some eigenvalues of a diagonal matrix after a symmetric rank-one update. Meyer [1970] discusses the use of a linesearch with Marquardt's method (see also Osborne [1972]). Shanno [1970] selects  $\lambda$  so that  $p_\lambda$  is a descent step for  $\|f(x)\|_2^2$ . The value  $\lambda = 0$  is tried first, and then increases are made by multiplying a threshold value by a factor greater than one until  $\psi'(\lambda) < 0$ , where  $\psi(\lambda) = \|f(x + p_\lambda)\|_2$ . In addition, a linesearch is also used when  $\cos(p_\lambda, g)$  is above a threshold value, that is, when  $p_\lambda$  is judged to be nearly in the direction of  $-g$ . Shanno's method is meant for general unconstrained or linearly-constrained minimization, as well as for nonlinear least squares.

Several methods have attempted to approximate Levenberg-Marquardt directions by a vector that is the sum of a component in the steepest descent direction, and a component in the Gauss-Newton direction  $p_{GN}$ . Jones [1970] combines searches along a spiral arc connecting  $p_{GN}$  and

the origin with parabolic interpolation in order to obtain a decrease in the sum of squares. If a reduction is not achieved after trying several arcs, then the steepest descent direction is searched. The method of Powell [1970a] for nonlinear equations and [1970b] for unconstrained optimization searches along a piecewise linear curve. The algorithm for unconstrained optimization requires some agreement between the reduction predicted by the quadratic model and the actual reduction in the sum of squares before the step is accepted. Global convergence results that include use of the quadratic model (2.1) for nonlinear least squares are given in Powell [1975] (see also Moré [1983]). Steen and Byrne [1973] approximate a search along an arc that intersects  $g$  at a nonzero point. Their algorithm requires that  $J^T J$  be scaled so that its smallest eigenvalue is 2, which they accomplish by computing  $(J^T J)^{-1}$  and finding either  $\|(J^T J)^{-1}\|_1$  or  $\|(J^T J)^{-1}\|_\infty$ . A diagonal of unspecified small magnitude is added to  $J^T J$  in the event of singularity. A difficulty with any algorithm based on this type of approach is that it is not clear how to define the approximation when the Gauss-Newton direction is not numerically well defined.

Fletcher [1971] implements a modified version of Marquardt's algorithm, in which adjustments in the parameter  $\lambda$  are made on the basis of a comparison of the actual reduction in the sum of squares

$$\frac{1}{2} \left( \|f(x + p_\lambda)\|_2^2 - \|f(x)\|_2^2 \right), \quad (3.7)$$

with the reduction

$$g^T p_\lambda + \frac{1}{2} p_\lambda^T J^T J p_\lambda \quad (3.8)$$

predicted by the model (3.2), which is the optimum value of the objective in (3.1) (see also Powell [1970b]). The step  $p_\lambda$  is taken only when there is sufficient agreement between (3.7) and (3.8), instead of accepting  $p_\lambda$  whenever the trial step results in a reduction in the sum of squares. Fletcher also introduces more complicated techniques for updating  $\lambda$ . The scheme for decreasing  $\lambda$  differs from that given by Marquardt in that division by a constant factor is used only until  $\lambda$  reaches a threshold value,  $\lambda_c$ , below which it is replaced by zero. This modification is motivated by a desire to allow the Gauss-Newton step ( $\lambda = 0$ ) when Gauss-Newton methods would work well, since  $\lambda$  is always positive in Marquardt's method, and to allow the initial choice of  $\lambda = 0$  rather than some arbitrary positive value. Because numerical experiments show that multiplying by a fixed constant factor may be inefficient, Fletcher uses safeguarded quadratic interpolation to increase  $\lambda$  when (3.7) and (3.8) differ substantially. If the current value of  $\lambda$  is nonzero, then it is divided by a factor

$$\gamma = \begin{cases} 0.1, & \text{if } \alpha_{min} < 0.1; \\ \alpha_{min}, & \text{if } \alpha_{min} \in [0.1, 0.5]; \\ 0.5, & \text{if } \alpha_{min} > 0.5. \end{cases} \quad (3.9)$$

where  $\alpha_{min}$  is the minimum of the quadratic interpolant to the function  $\phi(\alpha) = \|f(x + \alpha p)\|_2^2$  at  $\phi(0)$ ,  $\phi'(0)$ , and  $\phi(1)$ . There is also a provision to increase  $\lambda = 0$  to the threshold value  $\lambda_c$  under certain circumstances. The choice of  $\lambda_c$  appears to be a major difficulty.

Fletcher gives some theoretical justification for choosing  $\lambda_c$  to be the reciprocal of the smallest eigenvalue of  $(J^T J)^{-1}$ . Since he chooses to solve (3.3) directly for each value of  $\lambda$  via the Cholesky factorization, rather than compute the eigenvalue decomposition of  $J^T J$  or the singular values of  $J$ , the minimum eigenvalue of  $J^T J$  is not available without further computation. He therefore updates the estimate of  $\lambda_c$  only when  $\lambda$  is increased from 0, calculating  $(J^T J)^{-1}$  from the Cholesky factorization of  $J^T J$ , and then takes either  $\lambda_c = 1 / \|(J^T J)^{-1}\|_\infty$  or  $\lambda_c = 1 / \text{trace}((J^T J)^{-1})$ . A drawback is that  $\lambda_c$  is not defined when  $J^T J$  is singular, and it is not well defined when  $J^T J$  is ill-conditioned. Harwell subroutine VA07A is an implementation of Fletcher's method. It allows the user to select the scaling matrix  $D$ , which then remains fixed throughout the computation. The default for the scaling matrix is the square root of the diagonal of  $J^T J$  at the starting value.

An efficient and stable method for solving (3.3) for several values of  $\lambda$  based on the linear least-squares formulation (3.4) is given by Osborne [1972]. The method is accomplished in two stages. First, the  $QR$  factorization of  $J$  is computed, to obtain

$$\begin{pmatrix} Q & 0 \\ 0 & I \end{pmatrix} \begin{pmatrix} J \\ \sqrt{\lambda} D \end{pmatrix} = \begin{pmatrix} R \\ \sqrt{\lambda} D \end{pmatrix}, \quad (3.10)$$

after which a series of elementary orthogonal transformations are applied to reduce the right-hand side of (3.10) to triangular form. Thus it is only necessary to repeat the second stage of this procedure when the value of  $\lambda$  is changed, provided the  $QR$  factorization of  $J$  is saved. In a later paper, Osborne [1976] discusses a variant of Marquardt's algorithm for which he proves global convergence to a stationary point of  $f^T f$  under the assumption that the sequence  $\{\lambda_k\}$  remains bounded. In this method, he uses a simple scheme similar to the one proposed by Marquardt to update  $\lambda$ , but controls adjustments in  $\lambda$  by comparing (3.7) and (3.8). His implementation takes  $D$  to be the square root of the diagonal of  $J^T J$ , as in Marquardt's method.

The algorithm of Moré [1978] adjusts the step bound  $\delta$  in (3.1) rather than  $\lambda$ , a strategy used in trust-region methods for unconstrained optimization (see Moré [1983] for a survey). Changes in  $\delta$  depend on agreement between (3.7) and (3.8); increases are accomplished by taking  $\delta_{k+1} = 2\|D_k p_k\|_2$ , while  $\delta$  is decreased by multiplying by the factor  $\gamma$  defined by (3.9). In order to obtain  $\lambda$  when the bound in (3.1) is active, the nonlinear equation

$$\Psi(\lambda) = \|D p_\lambda\|_2 - \delta = \left\| (J^T J + \lambda D^T D)^{-1} g \right\|_2 - \delta = 0 \quad (3.11)$$



is approximately solved by truncating a safeguarded Newton method based on the work of Hebden [1973] (see also Reinsch [1971]). Moré reports that, on the average, (3.11) is solved fewer than two times per iteration. Also, he proves global convergence to a stationary point of  $f^T f$ , without assuming boundedness for  $\{\lambda_k\}$ . Many computational details are given, including an efficient method for calculating the derivative of  $\Psi(\lambda)$  in (3.11) that uses the  $QR$  factorization of  $J$ . A modification of the two-stage factorization described in Osborne [1972] that allows column pivoting is used to solve (3.3). Subroutine **LMDER** in MINPACK [Moré, Garbow, and Hillstom (1980)] is an implementation of the method. Variables are scaled internally in **LMDER** according to the following scheme: the initial scaling matrix  $D_0$  is the square root of the diagonal of  $J^T J$  evaluated at  $x_0$ , and the  $i$ th diagonal element of  $D_k$  is taken to be the maximum of the  $i$ th diagonal element of  $D_{k-1}$  and the square root of the  $i$ th diagonal element of  $J^T J$ . Numerical results are presented indicating that this scaling compares favorably with those used by Fletcher, and by Marquardt and Osborne. The user also has the option of providing an initial diagonal scaling matrix that is retained throughout the computation.

Nazareth [1980, 1983] describes a hybrid method that combines a Levenberg-Marquardt method with a quasi-Newton approximation  $H_k$  to the full Hessian. The search directions solve a system of the form

$$(\theta_k J_k^T J_k + (1 - \theta_k) H_k + \lambda_k D_k^T D_k) p = -g_k,$$

with  $\theta_k \in [0, 1]$  and  $\lambda_k \geq 0$ . He compares the reduction in the sum of squares predicted by both the Levenberg-Marquardt and quasi-Newton models with the actual reduction, and then chooses  $\theta_k$  on the basis of this comparison. In Nazareth [1983], a simple version of the hybrid strategy is implemented that uses Davidon's optimally conditioned update, with  $D_k = I$ , and a variation of Fletcher's [1971] method for updating  $\lambda$ . Results are reported for a set of eleven test problems — including five problems with nonzero residuals — and compared to the use of the algorithm as a quasi-Newton method ( $\theta_k = 0$ ) or a Levenberg-Marquardt method ( $\theta_k = 1$ ). He concludes that the hybrid method is somewhat better for the problems with nonzero residuals, and recommends development of a more sophisticated implementation.

#### 4. Corrected Gauss-Newton Methods

Gill and Murray [1976] propose a linesearch algorithm that divides  $\mathbb{R}^n$  into complementary subspaces  $\hat{\mathcal{R}}$  and  $\hat{\mathcal{N}}$ , where  $\hat{\mathcal{R}} \subseteq \mathcal{R}(J^T)$ , and  $\hat{\mathcal{N}}$  is nearly orthogonal to  $\mathcal{R}(J^T)$ . The search direction is the sum of a Gauss-Newton direction in  $\hat{\mathcal{R}}$ , and a projected Newton direction in  $\hat{\mathcal{N}}$ . This strategy avoids a shortcoming of Gauss-Newton methods — that components of the

search direction that are nearly orthogonal to  $\mathcal{R}(J^T)$  may not be well determined when  $J$  is ill-conditioned — because each component is computed from a reasonably well-conditioned subproblem. The vector  $x - x^*$  may become almost entirely in  $\mathcal{R}(J^T)$  in a Gauss-Newton method, yet the algorithm computes a search direction that is virtually orthogonal to  $\mathcal{R}(J^T)$  due to ill conditioning in the Jacobian (see Fraley [1987b]). Gill and Murray show that both Gauss-Newton algorithms defined by (2.4) and Levenberg-Marquardt algorithms generate search directions that lie in  $\mathcal{R}(J^T)$ , while the Newton search direction generally will have a component in  $\mathcal{N}(J)$ , the orthogonal complement of  $\mathcal{R}(J^T)$ , whenever  $J$  has linearly dependent columns. For problems with small residuals, they point out that  $J^T J$  is a reasonable approximation to the full Hessian in  $\mathcal{R}(J^T)$ , but not in  $\mathcal{N}(J)$ . Thus, in situations where  $x - x^*$  is orthogonal to  $\mathcal{R}(J^T)$ , and  $J$  is well-conditioned but has linearly dependent columns (for example, when  $m < n$ ), the Gauss-Newton and Levenberg-Marquardt directions have no component in the direction of  $x - x^*$ , while Newton's method and also the method of Gill and Murray would have components in both  $\mathcal{R}(J^T)$  and  $\mathcal{N}(J)$ .

The basic idea of the method is as follows. Suppose that

$$J = QTV^T \quad (4.1)$$

is an orthogonal factorization of  $J$ , in which  $T$  is triangular with diagonal elements in decreasing order of magnitude (either a  $QR$  factorization with column pivoting or the singular-value decomposition). Let

$$V = (Y \ Z) \quad (4.2)$$

be a partition of  $V$  into the first  $\text{gradc}(J)$  columns and the remaining  $n - \text{gradc}(J)$  columns. The columns of  $Y$  form an orthonormal basis for  $\hat{\mathcal{R}}$ , and those of  $Z$  form an orthonormal basis for  $\hat{\mathcal{N}}$ . The Newton search direction for the nonlinear least-squares problem is given by

$$(J^T J + B)p = -J^T f,$$

with

$$B = \sum_{i=1}^m \phi_i \nabla^2 \phi_i,$$

or, equivalently,

$$V^T(J^T J + B)p = -V^T J^T f, \quad (4.3)$$

since  $V$  is nonsingular. Using (4.2), equation (4.3) can be split into two equations:

$$Y^T(J^T J + B)p = -Y^T J^T f, \quad (4.4)$$

and

$$Z^T(J^T J + B)p = -Z^T J^T f. \quad (4.5)$$

Substituting  $p = Yp_y + Zp_z$  into (4.4) yields

$$Y^T J^T J Y p_y + Y^T J^T J Z p_z + Y^T B p = -Y^T J^T f.$$

Since  $\text{gradc}(J)$  is chosen to approximate  $\text{rank}(J)$ ,  $\|JZ\|$  is presumed to be zero, so that  $Y^T J^T J Z p_z$  vanishes. Also, for zero residual problems, the term  $Y^T B p$  would be small near a minimum relative to  $Y^T J^T J Y p_y$ , since  $\|B\|$  approaches zero. Defining  $\epsilon$  to be  $\|x - x^*\|$ , where  $x^*$  is a minimum at which the residuals are zero, and assuming  $\|f\| = O(\epsilon)$  we have

$$Y^T J^T J Y p_y = O(\epsilon); \quad Y^T B p = O(\epsilon^2); \quad Y^T J^T f = O(\epsilon).$$

The range-space component of the search direction is therefore chosen to satisfy

$$Y^T J^T J Y p_y = -Y^T J^T f. \quad (4.6)$$

With  $\text{gradc}(J) = \text{rank}(J)$ , the vector  $Y p_y$  is the minimal  $l_2$ -norm least-squares solution to  $Jp \approx -f$ , and is therefore a Gauss-Newton direction. For the null-space portion, since  $JZ = 0$  is assumed, (4.6) reduces to

$$Z^T B p = 0,$$

which may be solved for  $Z p_z$  given  $Y p_y$  from (4.5) using

$$Z^T B Z p_z = -Z^T B Y p_y. \quad (4.7)$$

When exact second derivatives are not available, the use of finite difference approximations along the columns of  $Z$  is suggested.

A version of this algorithm called the *corrected Gauss-Newton method* [Gill and Murray (1978)] forms the basis for the nonlinear least-squares software currently in the NAG Library [1984]. It uses the singular-value decomposition of  $J$ , rather than a  $QR$  factorization. Rules based on the relative size of the singular values are given for choosing an integer  $\text{gradc}(J)$  to approximate  $\text{rank}(J)$ , and an attempt is made to group together singular values that are similar in magnitude. The method is not as sensitive to  $\text{gradc}(J)$  as Gauss-Newton is to rank estimation, both because of the division of the computation of the search direction into separate components in  $\tilde{R}$  and  $\tilde{N}$ , and because  $\text{gradc}(J)$  is varied adaptively based on a measure of the progress of the minimization. Moreover, the rate of convergence is potentially faster than Gauss-Newton or Levenberg-Marquardt methods on problems with nonzero residuals. The quantity  $\text{gradc}(J)$  is reduced when the sum of squares is not adequately decreasing, so that there is the potential

of having  $\tilde{V} = \mathbb{R}^n$  (with exact second derivatives, this implies taking full Newton steps) in the vicinity of a solution. The derivation below shows how the corrected Gauss-Newton method differs from the earlier version based on the  $QR$  factorization.

Because of (4.1),  $J^T J$  can be written as  $V^T T^T T V$ , so that (4.3) is equivalent to

$$T^T T V^T p + V^T B p = -T^T Q^T f. \quad (4.8)$$

Using  $p = Y p_y + Z p_z$ , along with

$$V^T Y = \begin{pmatrix} I_{\text{grade}(J)} \\ 0 \end{pmatrix} \quad \text{and} \quad V^T Z = \begin{pmatrix} 0 \\ I_{n-\text{grade}(J)} \end{pmatrix},$$

(4.8) becomes

$$T^T T \begin{pmatrix} I_{\text{grade}(J)} \\ 0 \end{pmatrix} p_y + T^T T \begin{pmatrix} 0 \\ I_{n-\text{grade}(J)} \end{pmatrix} p_z + V^T B p = -T^T Q^T f. \quad (4.9)$$

If we let

$$T = \begin{pmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{pmatrix}$$

be a partition of  $T$ , where  $T_{11}$  is the submatrix consisting of the first  $k$  rows and columns of  $T$ , then

$$T^T T = \begin{pmatrix} (T_{11}^T T_{11} + T_{21}^T T_{21}) & (T_{11}^T T_{12} + T_{21}^T T_{22}) \\ (T_{12}^T T_{11} + T_{22}^T T_{21}) & (T_{12}^T T_{12} + T_{22}^T T_{22}) \end{pmatrix},$$

and (4.9) can be split into two equations :

$$(T_{11}^T T_{11} + T_{21}^T T_{21}) p_y + (T_{11}^T T_{12} + T_{21}^T T_{22}) p_z + Y^T B p = -(T_{11}^T \quad T_{12}^T) Q^T f, \quad (4.10)$$

and

$$(T_{12}^T T_{11} + T_{22}^T T_{21}) p_y + (T_{12}^T T_{12} + T_{22}^T T_{22}) p_z + Z^T B p = -(T_{12}^T \quad T_{22}^T) Q^T f. \quad (4.11)$$

As in the earlier version, the term  $Y^T B p$  is ignored in (4.10). Moreover, in the case that (4.1) is the singular-value decomposition, both  $T_{12}$  and  $T_{21}$  vanish and the two equations can be further simplified to

$$S_1^2 p_y = -(S_1 \quad 0) Q^T f, \quad (4.12)$$

and

$$S_2^2 p_z + Z^T B p = -(0 \quad S_2) Q^T f, \quad (4.13)$$

where

$$S_1 \equiv T_{11} \quad \text{and} \quad S_2 \equiv T_{22}.$$

Note that  $S_1$  and  $S_2$  are diagonal matrices, and that the  $p_y$  term in the second equation could not be ignored if (4.1) were a triangular factorization of  $J$ , because then  $(T_{12}^T T_{11} + T_{22}^T T_{21})$

could not be assumed negligible relative to  $(T_{12}^T T_{12} + T_{22}^T T_{22})$ . The equations that are ultimately solved are

$$S_1 p_y = -(I_{\text{grade}(J)} \quad 0) Q^T f, \quad (4.14)$$

and

$$(S_2^2 + Z^T B Z) p_x = -(0 \quad S_2) Q^T f - Z^T B p_y. \quad (4.15)$$

The matrix  $S_2^2 + Z^T B Z$  is replaced by a modified Cholesky factorization if it is computationally singular or indefinite. The range-space component is a Gauss-Newton search direction, while, in the positive-definite case, the null-space component is a projected Newton direction.

When no modification is necessary, the subproblem being solved is

$$\begin{aligned} \min_{p \in \mathbb{R}^n} g^T p + \frac{1}{2} p^T (J^T J + B) p \\ \text{subject to } Jp \simeq -f, \end{aligned} \quad (4.16)$$

where ' $\simeq$ ' is taken in a least-squares sense if the rows of  $J$  are linearly dependent, as in the case when  $m > n$ , and otherwise as equality. Subproblem (4.16) is an equality constrained quadratic program. When  $\text{rank}(J) = \text{grade}(J) = n$ , its solution is a full-rank Gauss-Newton direction that is completely determined by the constraints in (4.16). When  $\text{rank}(J) = \text{grade}(J) < n$ , the search direction is computed as the sum of two mutually orthogonal components, defined by equations (5.3.14) and (5.3.15). In this case  $S_2 = 0$ , so that the projected Hessian in (5.3.15) is  $Z^T B Z$  and therefore involves only the second derivatives of the residuals. We shall return to this point in Section 7, when we discuss SQP methods for nonlinear least squares.

Although the range-space component solving (4.14) can never be a direction of increase for  $f^T f$  (see Fraley [1987a]), the search direction computed by (4.14) and (4.15) may not be a descent direction for  $f^T f$ , regardless of whether or not  $S_2^2 + Z^T B Z$  is modified, on account of the  $p_y$  term in (4.15). Thus, if  $|\cos(g, p)|$  is smaller than some prescribed value, or if  $g^T p$  is positive, then a modified Newton search direction (corresponding to the case  $\text{grade}(J) = 0$ ) is used instead. A finite-difference approximation to the projected matrix  $Z^T B Z$  along the columns of  $Z$ , and a quasi-Newton approximation to  $B$  (see the discussion in Section 5) are given as alternatives to handle cases in which second derivatives of the residual functions are not available or are difficult to compute. Gill and Murray test their method on a set of twenty-three problems, and find that when quasi-Newton approximations to  $B$  are used, the algorithm does not perform as well as it does with exact second derivatives or finite-difference approximations to a projection of  $B$ . They observe only linear convergence for the quasi-Newton version on problems with large residuals. The algorithms are implemented in the NAG Library [1984] in the NAG Library [1984] subroutine E04HEF uses exact second derivatives, while subroutine E04GBF is the quasi-Newton version.

## 5. Special Quasi-Newton Methods

Another approach to the nonlinear least-squares problem is based on a quadratic model

$$g^T p + \frac{1}{2} p^T (J^T J + \bar{B}) p,$$

where  $\bar{B}$  involves quasi-Newton approximations to the term

$$B(x) = \sum_{i=1}^m \phi_i(x) \nabla^2 \phi_i(x)$$

in the Hessian of the nonlinear least-squares objective. Brown and Dennis [1971] first proposed a method in which the Hessian matrix of each of the residuals was updated separately. This technique is impractical because it entails the storage of  $m$  symmetric matrices of order  $n$ , and more recent research has aimed to approximate  $B$  as a sum.

Dennis [1973] suggests choosing the updates to satisfy a quasi-Newton condition

$$\bar{B}_{k+1} s_k = y_k - J_{k+1}^T J_{k+1} s_k, \quad (5.1)$$

where

$$s_k \equiv x_{k+1} - x_k \quad \text{and} \quad y_k \equiv g_{k+1} - g_k.$$

It is implied that the update can then be chosen as in the unconstrained case, although there is some ambiguity as to how this should be done. One possibility is to update  $\bar{B}_k$  directly to obtain  $\bar{B}_{k+1}$ , subject to a quasi-Newton condition such as (5.1) on  $\bar{B}_{k+1} s_k$ . Another approach consistent with Dennis' description is to modify  $\bar{H}_k = J_{k+1}^T J_{k+1} + \bar{B}_k$ , requiring the updated matrix  $\bar{H}_{k+1}$  to satisfy a quasi-Newton condition

$$\bar{H}_{k+1} s_k = y_k. \quad (5.2)$$

Then  $\bar{B}_{k+1} = \bar{H}_{k+1} - J_{k+1}^T J_{k+1}$  is the new approximation to  $B$  at  $x_{k+1}$ . Depending on the update and quasi-Newton conditions, the two alternatives may not yield the same result. Moreover, updates defined by minimizing the change in the inverse of  $\bar{B}_k$ , such as the BFGS update to  $\bar{B}_k$ , make no sense in this context, since the matrix  $B$  would not, by itself, be expected to be invertible.

Betts [1976] implements a linesearch method in which the symmetric rank-one update (see Dennis and Moré [1977]) is applied to  $\bar{B}$ , with the quasi-Newton condition

$$\bar{B}_{k+1} s_k = y_k - J_k^T J_k s_k. \quad (5.3)$$

This scheme is equivalent to applying the symmetric rank-one formula to the matrix  $\bar{H}_k = J_k^T J_k + \bar{B}_k$  with the updated matrix  $\bar{H}_{k+1}$  satisfying (5.2), and then taking  $\bar{B}_{k+1} = \bar{H}_{k+1} - J_{k+1}^T J_k$ . He compares this algorithm with a Gauss-Newton method, and also with a hybrid algorithm that starts with Gauss-Newton, switching to the augmented Hessian  $\bar{H}_k$  when the iterates are judged to be sufficiently close together to be near a solution. It is not clear whether the update is performed when  $\bar{B}$  is not used in the hybrid method. Betts reports observing quadratic convergence for the special quasi-Newton methods. For further discussion of these results, see Section 2.

Bartholomew-Biggs [1977] compares the PSB update (see Dennis and Moré [1977]) and the symmetric rank-one update applied directly to  $\bar{B}$  in a linesearch method. These updates are tested with the quasi-Newton condition (5.1), as well as with the condition

$$\bar{B}_{k+1} s_k = J_{k+1}^T f_{k+1} - J_k^T f_{k+1}, \quad (5.4)$$

which is derived from the relation

$$\begin{aligned} \sum_{i=1}^m \phi_i(x_{k+1}) \nabla^2 \phi_i(x_{k+1}) s_k &= \sum_{i=1}^m \phi_i(x_{k+1}) \left[ \nabla \phi_i(x_{k+1}) - \nabla \phi_i(x_k) + \mathcal{O}(\|s_k\|^2) \right] \\ &\approx J_{k+1}^T f_{k+1} - J_k^T f_{k+1} \end{aligned}$$

(see also Dennis [1976]). Bartholomew-Biggs points out that, in general, quasi-Newton approximations to  $B$  may not adequately reflect changes that are due to the contribution of the residuals. For example, when each residual function  $\phi_i$  is quadratic, and consequently each  $\nabla^2 \phi_i$  is constant,  $B_{k+1}$  may differ from  $B_k$  by a matrix of rank  $n$ . For this reason, he does some experiments with updating  $\tau \bar{B}_k$  for  $\tau = f_{k+1}^T f_k / f_k^T f_k$ , which is the appropriate scaling for the special case in which  $f_{k+1} = \tau f_k$  and the  $\phi_i$  are quadratic. In his implementation, a Levenberg-Marquardt step is used whenever the linesearch fails to produce an acceptable reduction in the sum of squares and  $\cos(g, p) > -10^{-4}$ . The scaled symmetric rank-one update with (5.4) is selected to compare with other methods after preliminary tests, because it exhibited the best overall performance, and required fewer Levenberg-Marquardt steps. The other methods tested include a Gauss-Newton method, a method that combines Gauss-Newton with a Levenberg-Marquardt method, an implementation of Fletcher's [1971] Levenberg-Marquardt method, and a quasi-Newton method for unconstrained optimization. All of the fourteen test problems have nonzero residuals. Bartholomew-Biggs finds that the special quasi-Newton method is more robust than the other specialized methods for nonlinear least-squares, and that it is particularly suitable for problems with large residuals. He also observes that on problems on which the Gauss-Newton and Levenberg-Marquardt-based methods perform poorly, the special quasi-Newton method is more effective than the quasi-Newton method for general unconstrained optimization. Nothing

is said about the observed rate of convergence for any of the methods. He concludes that further research is needed to determine the best updating strategy, some desirable features being hereditary positive definiteness, and the ability to update a factorization of  $\tilde{B}$ . Finally, he indicates that it would be worthwhile to develop a hybrid method combining Gauss-Newton with a special quasi-Newton method, in order to avoid the cost of the updates on problems that are easily solved by Gauss-Newton methods.

Gill and Murray [1978] discuss a linesearch method in which they use the augmented Gauss-Newton quadratic model only to compute a component of the search direction in a subspace that approximates the null space of the Jacobian (see the preceding section). They apply the BFGS formula for unconstrained optimization (see Dennis and Moré [1977]) to the matrix  $\tilde{H}_k = J_{k+1}^T J_{k+1} + \tilde{B}_k$  with the quasi-Newton condition (5.2), and then form  $\tilde{B}_{k+1} = \tilde{H}_{k+1} - J_{k+1}^T J_{k+1}$ . The choice of the BFGS update is based on performance comparisons to a number of other updates, including the symmetric rank-one update and Davidon's optimally-conditioned update [Davidon (1975)], as well as the symmetric rank-one update applied to  $\tilde{H}_k = J_k^T J_k + \tilde{B}_k$  used in Betts [1976]. They point out that, if  $J_{k+1}^T J_{k+1} + \tilde{B}_k$  is positive definite, and  $y_k^T s_k > 0$ , then  $J_{k+1}^T J_{k+1} + \tilde{B}_{k+1}$  is also positive definite with this scheme. In order to safeguard the method, the projected approximate Hessian is replaced by a modified Cholesky factorization when it is singular or indefinite. In addition, if  $\cos(p, g)$  exceeds a fixed threshold value, a modified Newton step with the full augmented approximate Hessian is taken. See Section 4 for a summary of their observations on the performance of the methods.

Dennis, Gay, and Welsch [1981a] apply a scaled DFP update (see Dennis and Moré [1977]) to  $\tilde{B}_k$  at each step. The new approximation  $\tilde{B}_{k+1}$  solves

$$\min_{B, H} \|H^{-1/2}(\tau_k \tilde{B}_k - B)H^{-1/2}\|_F \quad (5.5)$$

subject to

$$H s_k = y_k : \quad H \text{ positive definite} \quad (5.6)$$

$$B s_k = J_{k+1}^T f_{k+1} - J_k^T f_{k+1} : \quad B \text{ symmetric.} \quad (5.7)$$

where

$$\tau_k \equiv \min\{|y_k^T s_k / s_k^T \tilde{B}_k s_k|, 1\}. \quad (5.8)$$

The scale factor  $\tau_k$  is based on the observation that the quasi-Newton approximation to  $B$  is often too large with the unscaled update, on account of the contribution of the residuals. The term  $|y_k^T s_k / s_k^T \tilde{B}_k s_k|$  in  $\tau_k$  is derived from the self-scaling principles for quasi-Newton methods of Oren [1973], and attempts to shift the eigenvalues of the approximation  $\tilde{B}_k$  to overlap with those of  $B_k$ , using new curvature information at  $x_k$ . This method forms the basis for the ACM



computer program NL2SOL [Dennis, Gay, and Welsch (1981b)], which is distributed by the PORT Library [1984] as subroutines N2G and DN2G. It is implemented as an adaptive method, in that Gauss-Newton steps are taken if the Gauss-Newton quadratic model predicts the reduction in the function better than the quadratic model that includes the term involving  $\tilde{B}$ . A trust-region strategy is used to enforce global convergence. Numerical results are given in Dennis, Gay, and Welsch [1981a] for a set of twenty-four test problems, many with two or three different starting values. medskip

Al-Baali and Fletcher [1985] describe some linesearch methods that are similar to the method of Dennis, Gay, and Welsch [1981a] discussed above. They observe that the DFP update defined by (5.5) - (5.8) is equivalent to finding  $\tilde{H}_{k+1}$  to solve

$$\min_{\tilde{H}, H} \|H^{-1/2}(J_{k+1}^T J_{k+1} + \tau_k \tilde{B}_k - \tilde{H})H^{-1/2}\|_F \quad (5.9)$$

subject to

$$H s_k = y_k; \quad H \text{ positive definite} \quad (5.10)$$

$$\tilde{H} s_k = J_{k+1}^T f_{k+1} - J_k^T f_{k+1} + J_{k+1}^T J_{k+1} s_k; \quad \tilde{H} \text{ symmetric,}$$

where

$$\tau_k \equiv \min\{|y_k^T s_k / s_k^T \tilde{B}_k s_k|, 1\},$$

and then forming

$$\tilde{B}_{k+1} = \tilde{H}_{k+1} - J_{k+1}^T J_{k+1}.$$

Moreover, they use the condition

$$H s_k = \tilde{y}_k; \quad H \text{ positive definite,} \quad (5.11)$$

with

$$\tilde{y}_k \equiv J_{k+1}^T J_{k+1} s_k + J_{k+1}^T f_{k+1} - J_k^T f_{k+1} = y_k + \mathcal{O}(\|s_k\|^2) \quad (5.12)$$

as an alternative to (5.10), and mention that (5.10) has been replaced by (5.11) in newer versions of NL2SOL. The claim is that the updated matrix is almost always positive definite. However, if the matrix  $J_{k+1}^T J_{k+1} + \tau_k \tilde{B}_k$  is not positive semi-definite,  $\tau_k$  is replaced by a quantity  $\hat{\tau}_k$  that is calculated by a method similar to a Rayleigh quotient iteration, so that  $J_{k+1}^T J_{k+1} + \hat{\tau}_k \tilde{B}_k$  is positive semi-definite and singular. A corresponding BFGS method is also given in which the update is defined by

$$\min_{\tilde{H}, H} \|H^{-1/2}((J_{k+1}^T J_{k+1} + \tau_k \tilde{B}_k)^{-1} - \tilde{H}^{-1})H^{-1/2}\|_F$$

instead of (5.9). They conclude from computational tests (described in Al-Baali [1984]) that their method is somewhat more efficient in terms of the number of Jacobian evaluations than NL2SOL,

but requires more function evaluations, and that there is no significant difference between the DFP and BFGS updates. Al-Baali and Fletcher also introduce scaling factors based on finding a measure of the error in the inverse Hessian. They observe that, for the BFGS update for unconstrained optimization,

$$\|H_k^{-1/2}(H_k^{-1} - H_{k+1}^{-1})H_k^{-1/2}\|_F^2 = \Delta_k(H_k; y_k),$$

where

$$\Delta_k(H_k; y_k) \equiv \left( \frac{y_k^T H_k^{-1} y_k}{y_k^T s_k} \right)^2 - 2 \frac{y_k^T s_k}{s_k^T H_k s_k} + 1. \quad (5.13)$$

Hence an "optimal" value of  $\tau$  can be found by minimizing  $\Delta_k(J_{k+1}^T J_{k+1} + \tau \tilde{B}_k)$  as a function of  $\tau$ . Newton's method is used to find  $\tau$ , an iterative process that requires factorization of  $J_{k+1}^T J_{k+1} + \tau \tilde{B}_k$  for each intermediate value of  $\tau$ . They were apparently unable to draw any broad conclusions from numerical experiments with this scaling, and refer to Al-Baali [1984] for details.

A convergence analysis for minimization algorithms based on a quadratic model in which part of the Hessian is computed by a quasi-Newton method is given by Dennis and Walker [1981] (see also Chapter 11 of Dennis and Schnabel [1983]). These results are restricted to methods that satisfy a least-change condition on the matrix  $\tilde{B}_k$  (analogous to the PSB and DFP updates). Only a fairly mild assumption is needed to prove superlinear convergence to an isolated local minimum  $x^*$ : that the vector  $y_k^B$  in the quasi-Newton condition

$$\tilde{B}_k s_k = y_k^B$$

be chosen so that the norm of the update is

$$\mathcal{O}(\max\{\|x_k - x^*\|^p, \|x_{k+1} - x^*\|^p\}),$$

for some  $p > 0$ . This assumption is satisfied for  $y_k^B$  in each quasi-Newton update to  $\tilde{B}_k$  described above. Their treatment of inverse updates is for the case in which part of the inverse Hessian is computed, and hence does not apply here. To the best of our knowledge, no convergence results have yet been proven for scaled versions of the updates, or for updates to  $J_{k+1}^T J_{k+1} + \tilde{B}_k$  that are not equivalent to some direct quasi-Newton update to  $\tilde{B}_k$ .

## 6. Conjugate-Gradient Acceleration of Gauss-Newton Methods

Ruhe [1979] uses preconditioned conjugate gradients to speed up convergence of Gauss-Newton methods. General references on conjugate gradients include Fletcher [1980], Chapter 4, and Gill, Murray, and Wright [1981], Chapter 4. We give a brief explanation below.

The linear conjugate gradient method minimizes an  $n$ -variate quadratic function

$$Q(x) = q^T x + \frac{1}{2} x^T H x,$$

in at most  $n$  iterations. The iteration is

$$p_k = -g_k + \beta_{k-1} p_{k-1}; \quad (6.1)$$

$$x_{k+1} = x_k + \alpha_k p_k$$

where

$$\alpha_k = \frac{\|g_k\|_2^2}{p_k^T H p_k}; \quad \beta_k = \frac{\|g_{k+1}\|_2^2}{\|g_k\|_2^2};$$

$$g_k = \nabla Q(x_k) = q + H x_{k+1}.$$

The method produces a sequence of search directions that are  $H$ -conjugate, that is

$$p_i^T H p_j = 0 \quad \text{if } i \neq j.$$

The number of iterations needed to minimize  $Q$  by conjugate gradients (with exact arithmetic) is equal to the number of distinct eigenvalues of  $H$ . The idea of preconditioning is to transform  $H$  into a matrix whose eigenvalues are nearly identical in magnitude. If a positive-definite matrix  $W$  is used as a *preconditioner*, then convergence occurs in the same number of steps that would be taken for a quadratic function with the Hessian matrix

$$W^{-1/2} H W^{-1/2}.$$

The ideal preconditioner would be  $W = H$ , but since conjugate gradients are competitive mainly when  $n$  is large, an approximation that is relatively inexpensive to factorize is used. For a smooth nonlinear function  $\mathcal{F}(x)$ , the conjugate gradient method (6.1) can also be applied, with  $g_k = \nabla \mathcal{F}(x_k)$  and  $\alpha_k$  determined by a linesearch, with safeguards to ensure descent. There are several possible choices for  $\beta_k$  that are equivalent to the one given above for the quadratic case (see, for example, Fletcher [1981], Chapter 4). The method is often restarted every  $n$  iterations on account of the variation in  $\nabla^2 \mathcal{F}(x)$  for non-quadratic functions (e. g., Gill, Murray, and Wright [1981], Chapter 4). Preconditioners for the non-quadratic case attempt to approximate  $\nabla^2 \mathcal{F}(x)$ .

In Ruhe's algorithm, the matrix  $J^T J$  is used as the preconditioner, and an orthogonal factorization of  $J$  is used to compute the necessary quantities. The method is applied to problems in which the residuals are nonzero and the Jacobian has full rank, and is restarted every  $n$  iterations. He concludes that the preconditioned conjugate-gradient method never increases the total number of iterations required to solve a given problem relative to Gauss-Newton, and that

significant improvements in the speed of linear convergence of Gauss-Newton on large-residual problems can be achieved with conjugate-gradient acceleration.

Al-Baali and Fletcher [1985] point out that conjugate-gradient acceleration of the type described by Ruhe is equivalent to applying a BFGS update to the Gauss-Newton approximate Hessian  $J^T J$  at each step. They implement and test both this method (without restarts) and a scaled version, where the scale parameter  $\tau$  is chosen to minimize  $\Delta_k(\tau J_k^T J_k; \tilde{y}_k)$  as a function of  $\tau$  (see (5.4.14)). They give no conclusions as to the relative efficiency of the scaled and unscaled versions of the method, but find that the modified methods offer some improvement over Gauss-Newton, while exhibiting the same difficulties.

## 7. Sequential Quadratic Programming (SQP) Methods

Fraley [1987a] proposes algorithms that solve quadratic programming subproblems whose formulation is based on convergence properties of sequential quadratic programming methods for constrained optimization, and on geometric considerations in nonlinear least squares. The motivation behind these methods is as follows. Recall that the Hessian matrix of the least-squares objective can be separated into the sum of two components involving different types of derivative information:

$$\nabla^2 \left( \frac{1}{2} f^T f \right) = J^T J + B,$$

where

$$B \equiv \sum_{i=1}^m \phi_i \nabla^2 \phi_i.$$

The corrected Gauss-Newton methods (Section 4) calculate a search direction that is separated into two orthogonal components when  $0 < \text{gradc}(J) < n$ , and can be viewed as SQP methods. When  $\text{gradc}(J) = \text{rank}(J) < n$ , the contributions of  $J^T J$  and of  $B$  (or of an approximation to  $B$ ) are essentially decoupled because the contribution of  $J^T J$  in the projected Hessian is zero. No such separation is possible when  $\text{rank}(J) = n$ . In any case,  $\text{gradc}(J) < n$  may be selected based on the progress of the minimization as well as the singular values of  $J$ , so that partial separation of  $J^T J$  and  $B$  may occur between the extremes of Gauss-Newton ( $\text{gradc}(J) = \text{rank}(J)$ ), and a full Newton-type method ( $\text{gradc}(J) = 0$ ). The strategy of making a quasi-Newton approximation to  $B$  which is then added to  $J^T J$  in a full Newton-type method has not been successful outside a neighborhood of the solution, unless it is combined with other techniques (see Section 5). The approach taken in Fraley [1987a] is to use a quasi-Newton approximation to the full Hessian, while separating out some of the contribution to the curvature due to  $J^T J$  by including first-order information about the residuals as constraints.

A search direction is computed as the solution to a quadratic program (QP) of the form

$$\min_{p \in \mathbb{R}^n} g^T p + \frac{1}{2} p^T H p \quad (7.1)$$

subject to

$$-b^L \leq Ap + c \leq b^U$$

where

$$b^L \geq 0 \quad \text{and} \quad b^U \geq 0.$$

In SQP methods for constrained optimization,  $H$  approximates the Hessian of a Lagrangian function in order to take into account the curvature of the constraints that are active at the solution (e. g., Powell [1983], Gill et al. [1985b, 1986b], Nocedal and Overton [1985], Stoer [1985], and Gurwitz [1986]). For nonlinear least squares, it suffices for  $H$  to approximate the Hessian matrix of  $\frac{1}{2} f^T f$  even if some of the constraints in (7.1) are active at a solution  $x^*$ , because  $g(x^*) = 0$ . These methods have the potential to converge faster than quasi-Newton methods for unconstrained optimization, since only the projection of the Hessian in the null space of the active QP constraint normals — rather than the full Hessian — need be positive definite as a condition for superlinear convergence.

Two classes of suitable QP constraints for (7.1) are described: constraints on the directional derivatives of individual residuals, and constraints based the  $QR$  factorization of  $J$ . A departure from other algorithms is that information about the residuals, and interrelationships between residuals, can be used to construct the subproblems (the algorithm of Davidon [1976] is an exception — see Section 11). In the SQP algorithms, a set  $C$  of desirable constraints is chosen first, which may be infeasible or may otherwise exclude all suitable search directions. For example, such a set of constraints is

$$\{\nabla \phi_i^T p = \phi_i; \quad i = 1, 2, \dots, m\}. \quad (7.2)$$

Any  $p$  satisfying  $\nabla \phi_i^T p = -\phi_i$  is a descent direction for  $\phi_i$  if  $\phi_i \neq 0$  and is otherwise orthogonal to  $\nabla \phi_i$ . The unconstrained minimum  $p_{QN}$  of the QP objective in (7.1) is a descent direction for the nonlinear least-squares objective provided  $H$  is positive definite. Therefore, as long as  $p_{QN}$  is considered satisfactory, an acceptable search direction will eventually be obtained by either removing some constraints from  $C$ , or else by perturbing the constraints in  $C$  so as to enlarge the feasible region. Based on this reasoning, she proposes two different strategies (which could also be combined)

One strategy uses a QP to select a subset of constraints in  $C$  as the feasible region for (7.1). Several quadratic programs may be solved within a single iteration in order to compute a search direction, which is justified for two reasons. First, starting the solution process for a QP with information about the solution of a related subproblem can often lead to significant savings in QP iterations (see, e. g., Gill et al. [1985a]). Also, when the cost of a function evaluation is much greater than the cost of a QP iteration, the effort involved in obtaining the search direction by solving more than one subproblem may be worthwhile if it results in a substantial reduction in the number of outer iterations.

It is difficult to automate the selection of QP constraints, and the evaluation of the current QP solution as a candidate for the search direction. One strategy considers each of the constraints in (7.2) separately in order of decreasing residual size, with the object of including as many of the constraints as possible. A constraint is added to the current constraint set (initially empty) if the corresponding QP computes an "acceptable" search direction  $\bar{p}$ . In addition to the requirement that  $g^T \bar{p} < 0$ , Fraley uses a lower bound on the magnitude of  $\bar{p}$ , and an upper bound on  $|\cos(g, \bar{p})|$ , as the criteria for accepting  $\bar{p}$ . Some other examples that use constraints based on the  $QR$  factorization are very similar to corrected Gauss-Newton methods (Section 4).

In the second approach, constraints in  $C$  are modified in order to obtain a suitable feasible region. This is accomplished by treating constraint bounds as variables in a QP. Using the constraint set (7.2), Fraley shows how these SQP algorithms are related to Gauss-Newton and Levenberg-Marquardt methods. The QP

$$\begin{aligned} & \min_{b, p} b^T b \\ & \text{subject to} \\ & -b \leq Jp + f \leq b \\ & b \geq 0, \end{aligned} \tag{7.3}$$

computes the smallest possible perturbation that allows all of the (7.2) to intersect. In the solution  $(\bar{b}; \bar{p})$  to (7.3), the vector  $\bar{p}$  is a Gauss-Newton search direction. When  $J$  is ill-conditioned, it is possible that the constraints in (7.2) do intersect ( $b = 0$ ), but that the intersection occurs at a vector  $\bar{p}$  that is very large in magnitude. For  $\omega > 0$ , the QP

$$\begin{aligned} & \min_{b, p} b^T b + \omega p^T p \\ & \text{subject to} \\ & -b \leq Jp + f \leq b \end{aligned} \tag{7.4}$$

$$b \geq 0,$$

forces  $\|h\|$  to increase when  $\|p\|$  would otherwise be large. In the solution  $(\hat{b}; \hat{p})$  to (7.4), the vector  $\hat{p}$  is a Levenberg-Marquardt search direction. In an SQP algorithm based on (7.3) (respectively, (7.4)) there is the option of using  $\hat{p}$  ( $\hat{p}$ ) as a search direction, or of using  $\hat{b}$  ( $\hat{b}$ ) to define bounds for a second QP of the form (7.1), from which the search direction is computed.

Fraley proposes a number of variations of these basic SQP algorithms and tests some of them on a set of fourteen problems. She uses the BFGS method to approximate  $H$  in (7.1) just as in unconstrained optimization, and observes that the approximation retains positive definiteness throughout. She finds the SQP methods work well on some problems, and poorly on some others, so that it is not possible to say anything conclusive about their performance relative to existing methods.

## 8. Continuation Methods

Continuation methods have also been applied to nonlinear least-squares problems. These methods solve a sequence of parameterized subproblems

$$\min \Phi(x; \tau_i); \quad i = 1, 2, \dots, i_{\max} \quad (8.1)$$

where

$$0 = \tau_0 < \tau_1 < \dots < \tau_{i_{\max}} = 1$$

and

$$\arg \min \Phi(x; 0) = x_0 \quad \text{and} \quad \arg \min \Phi(x; 1) = x^*.$$

The idea is that methods that have fast local convergence, but may not be robust in a global sense, can be applied to solve each subproblem in relatively few steps, because information from the solution of previous subproblems may be used to predict a good starting value for the next one.

DeVilliers and Glasser [1981] define

$$\Phi(x; \tau) \equiv \frac{1}{2} \|f(x)\|_2^2 + \frac{1}{2} (\tau^k - 1) \|f(x_0)\|_2^2 \quad (8.2)$$

where  $k$  is a positive integer, with a fixed spacing between the parameters  $\tau_i$  in (8.1). They test two different continuation methods, one that uses Newton's method (with linesearch) to solve the intermediate problems, and one that uses a Gauss-Newton method (with linesearch). An unspecified "device" is included in the implementation of both minimization techniques to ensure a decrease in the objective at every iteration. The continuation methods are compared with

results obtained by applying both minimization algorithms to the original problem. Intermediate subproblems are not solved exactly; the criterion

$$\|\nabla_x \Phi(x; \tau_i)\|_2 \leq \epsilon_i,$$

where  $\epsilon_i = 10^{-2}$  if  $i < i_{\max}$ , and  $\epsilon_{i_{\max}} = 10^{-6}$ , is used to determine convergence of a subproblem.

Numerical experiments are carried out on three different test problems, with multiple starting values, most of which are points of failure for both Newton's method and Gauss-Newton. They conclude that, although the continuation method is less efficient than the underlying method when both are successful, it will converge on many problems for which the underlying method fails when used alone. However, the results they present are for different values of the step size, and the exponent  $k$ , and no mechanism is given for the automatic choice of either of the parameters. DeVilliers and Glasser point out that their methods may require modification if the optimization method that is used to solve the subproblems encounters difficulties, or if the continuation path is not well-behaved. Fraley [1987a, 1988] observes that the first two test problems of DeVilliers and Glasser are very sensitive to the choice of the maximum step bound, or the initial trust-region size for most methods and that the methods can be quite efficient provided an appropriate non-default choice is made for these parameters.

Salane [1987] incorporates a trust-region strategy into a continuation method by defining

$$\Phi(x; \tau) \equiv \frac{1}{2} \left( \|f(x)\|_2^2 + (\tau - 1)\|f(x_0)\|_2^2 + \lambda(\tau - 1)\|D(x - x_0)\|_2^2 \right), \quad (8.3)$$

and then applying Gauss-Newton to this function for the inner iterations. Instead of allowing the continuation parameter  $\tau$  to range from 0 to 1, he advocates stopping when it becomes inefficient to solve the subproblems, and then restarting the method after replacing  $x_0$  by the new iterate. He points out that his approach is especially suitable for large-residual problems, because it transforms the original problem into a sequence of subproblems with small residuals. The idea is to attempt to determine when the neglected terms become significant, and then pose a new subproblem. An initial value,  $\tau_1$ , of the continuation parameter must be supplied by the user in order to start the method. Should any step fail to obtain a decrease in either the nonlinear least-squares objective or its gradient,  $\tau_1$  is decreased, and the calculation is repeated without changing  $x_0$ . Theorems on descent conditions and convergence are presented. Salane argues that his continuation method allows direct selection of the Levenberg-Marquardt parameter  $\lambda$  in (8.3), because  $\lambda$  may be chosen so that the term  $\lambda(1 - \tau)D^T D$  behaves somewhat like the second-order terms that have been neglected in the Hessian of  $\Phi(x; \tau)$ . However, no mechanism is suggested for automatic choice of  $\lambda$ , and  $\lambda = \|f(x_0)\|_2$  is used in the tests.



Salane gives test results for a version of his algorithm on a set of nine problems (all of which are included in our set). A comparison is made to results obtained from MINPACK, and also to the results reported by DeVilliers and Glasser [1981] for two of the test problems. He concludes that the performance of the method compares favorably with that of MINPACK, and is superior to the DeVilliers and Glasser continuation method on the relevant problems. The matrix  $D$  in (8.3) is taken to be the identity matrix throughout the tests, and for one test problem a type of variable scaling is used. No information is given concerning scaling for the MINPACK tests. The results that are presented correspond to several different values of  $\tau_1$ , although the criterion used in choosing this value is not given. Test results in which the value of  $\tau_1$  is varied are included for three of the problems for the purpose of showing that performance is sensitive to the specification of the continuation parameter.

## 9. Modifications of Unconstrained Optimization Methods

Besides Gauss-Newton methods, several straightforward modifications of unconstrained optimization methods are possible for nonlinear least squares. In quasi-Newton methods,  $J_0^T J_0$  can be used as the initial approximation to the Hessian matrix. Ramsin and Wedin [1977] report favorable results with this technique. We note that a perturbed matrix  $\tilde{J}_0^T \tilde{J}_0$  can be used as the initial approximate Hessian, where  $\tilde{J}_0$  is a modified Cholesky factor of  $J_0^T J_0$  (Gill and Murray [1974]), in order to maintain positive definiteness when  $J_0$  is ill-conditioned.

Wedin [1974] (see also Ramsin and Wedin [1977]) suggests a modification of Newton's method in which the search direction is defined by

$$(J^T J + \sum_{i=1}^m \tilde{\phi}_i \nabla^2 \phi_i) p = -g, \quad (9.1)$$

where  $\tilde{\phi}_i$  is the  $i$ th component of the projection  $\tilde{f}$  of  $f$  onto  $\mathcal{R}(J)$ . This iteration approaches Newton's method in the limit, since  $f(x^*) = \tilde{f}(x^*)$ , and is parameter-independent, in the sense that minimization of  $f$  as a function of  $x$  is equivalent to minimization of  $f$  as a function of a new variable  $z$  — provided the mapping that defines  $x$  as a function of  $z$  has a nonsingular Jacobian. An obvious difficulty is that  $\tilde{f}$ , and hence (9.1), is not well-defined when  $J$  is ill-conditioned.

Recall that in quasi-Newton methods for unconstrained optimization, the approximate Hessian matrix is required to satisfy the condition

$$H_k s_k = y_k, \quad (9.2)$$

where

$$s_k \equiv x_{k+1} - x_k \quad \text{and} \quad y_k \equiv g_{k+1} - g_k$$

(e.g. Dennis and Moré [1977]). Al-Baali and Fletcher [1985] suggest the use of  $\bar{y}_k$  defined by (5.12) rather than  $y_k = g_{k+1} - g_k$  in the quasi-Newton condition (9.2). They report improvements with the BFGS and DFP formulas when this substitution is made. However, they remark that the condition  $\bar{y}_k^T s_k > 0$  for hereditary positive definiteness of the updates is not guaranteed by the linesearch requirements, and they replace  $\bar{y}_k^T s_k$  in the update formulas by  $\max \{\bar{y}_k^T s_k, 0.01 y_k^T s_k\}$  as a safeguard. They do not consider this a major drawback, because  $\bar{y}_k^T s_k > 0.01 y_k^T s_k$  almost always occurred in their examples. A somewhat different safeguard is used in a later related paper [see Fletcher and Xu (1986), p. 26] discussed below.

Al-Baali and Fletcher [1985] also develop several hybrid linesearch methods in which the models are assessed in terms of the function  $\Delta_k$  defined by (5.13), an approximate measure of the error in the inverse Hessian. In one class of methods, the modified BFGS update described in the preceding paragraph is applied to a matrix of the form

$$H_{k+1} = (1 - \theta_k) H_k + \theta_k \tau_k J_{k+1}^T J_{k+1},$$

where  $\tau_k$  minimizes  $\Delta_k(\tau_k J_{k+1}^T J_{k+1}; \bar{y}_k)$ , and  $\theta_k$  is chosen to minimize  $\Delta_k(H_{k+1}; \bar{y}_k)$ , in order to obtain the new approximate Hessian. In their implementation, in which  $\theta_k$  is restricted to be either 0 or 1, they find that the method has difficulties on singular problems, and that the scaling of the search direction often does not allow  $\alpha = 1$  as a trial step in the linesearch. They refer to Al-Baali [1984] for more details of the tests.

Another class of hybrid methods defined by Al-Baali and Fletcher compares the value

$$\Delta_{QN} \equiv \Delta_k(H_k; \bar{y}_k)$$

for the current quasi-Newton approximation  $H_k$  with

$$\Delta_{GN} \equiv \Delta_k(J_{k+1}^T J_{k+1}; \bar{y}_k)$$

for the Gauss-Newton approximation. The basic algorithm can be summarized as follows :

$$\begin{aligned} \text{if } \Delta_{QN} < \Delta_{GN} & \text{ then use the modified BFGS search direction} \\ & \text{else use the Gauss-Newton search direction} \end{aligned} \quad (9.3)$$

They test several versions of this method that differ in the action taken whenever a switch from Gauss-Newton to quasi-Newton takes place. In one,  $H_{k+1}$  is reset to  $J_{k+1}^T J_{k+1}$ , while in another  $H_{k+1}$  is reset to the result of applying the modified BFGS update to  $J_{k+1}^T J_{k+1}$  (conjugate-gradient acceleration). They observe little difference in performance between these two alternatives, and find them to be the best of the many methods for nonlinear least squares treated in their study. A version of the first strategy that substitutes the quantity  $\min_{\tau} \Delta_k(\tau J_{k+1}^T J_{k+1}; \bar{y}_k)$

for  $\Delta_{GN}$  in the comparison with  $\Delta_{QN}$  is also tried, but it is found to have some difficulties on a problem for which the Jacobian is singular at the solution. A final variant maintains the quasi-Newton update throughout, and never resets the approximate Hessian. They find that this method is not as efficient as the others on some types of large-residual problem.

Fletcher and Xu [1986] give an example in which the hybrid method (9.3) has a linear rate of convergence when the BFGS method would converge superlinearly. The difficulty is that the comparison between  $\Delta_{QN}$  and  $\Delta_{GN}$  may fail to distinguish between zero-residual problems and those with nonzero residuals. They propose two new hybrid algorithms and show them to be superlinearly convergent. The first algorithm computes the modified BFGS search direction if

$$\frac{\|f(x_k)\|_2 - \|f(x_{k+1})\|_2}{\|f(x_k)\|_2} < \sigma, \quad (9.4)$$

for some fixed  $\sigma \in (0, 1)$ , and a Gauss-Newton step otherwise. The method is motivated by the following relationship

$$\lim_{k \rightarrow \infty} \frac{\|f(x_k)\|_2 - \|f(x_{k+1})\|_2}{\|f(x_k)\|_2} = \begin{cases} 0, & \text{if } \|f(x^*)\|_2 \neq 0; \\ 1, & \text{if } \|f(x^*)\|_2 = 0. \end{cases}$$

The second algorithm computes a modified BFGS step if

$$\frac{\|f(x_k) - f(x_{k+1})\|_2}{\|f(x_k)\|_2} < \sigma \quad \text{and} \quad \frac{\Delta_k(J_{k+1}^T J_{k+1}; \bar{y}_k)}{\Delta_k(J_k^T J_k; \bar{y}_k)} \geq \gamma, \quad (9.5)$$

where both  $\sigma$  and  $\gamma$  are fixed parameters in  $(0, 1)$ , and a Gauss-Newton step otherwise. The additional condition for choosing the BFGS search direction is derived from another asymptotic relationship

$$\lim_{k \rightarrow \infty} \frac{\Delta_k(J_{k+1}^T J_{k+1}; \bar{y}_k)}{\Delta_k(J_k^T J_k; \bar{y}_k)} = \begin{cases} 0, & \text{if } \|f(x^*)\|_2 = 0; \\ 1, & \text{if } \|f(x^*)\|_2 \neq 0. \end{cases}$$

Numerical results are given for set of fifty-six test problems, a few with multiple starting values. They conclude that the new methods offer some overall improvement over those based on (9.3), but that there is no reason to prefer the more complicated test (9.5) over (9.4).

## 10. Special Linesearches

Lindström and Wedin [1984] and Al-Baali and Fletcher [1986] propose specialized linesearch methods for nonlinear least-squares problems in which each residual is interpolated by a quadratic function, in contrast to the strategy of interpolating to the sum of squares used in conventional linesearches for unconstrained minimization. As a result a quartic polynomial, rather than a simpler cubic or quadratic, is minimized at each iteration of the linesearch.

Lindström and Wedin substitute their linesearch, which uses only function values, for the quadratic interpolation and cubic interpolation routines in the NAG Library (1980 version) nonlinear least-squares algorithm E04GBF (see Sections 4 and 5), and compare the performance with the NAG linesearch routines on a set of eighteen test problems. They find that no linesearch algorithm is superior over all, but that their algorithm makes a better initial prediction to the steplength that minimizes the sum of squares along the search direction. In a second set of tests that includes multiple starting values for many of the test problems, they add a modified version of their linesearch algorithm that reverts to a simple backtracking strategy if an acceptable decrease in the sum of squares is not obtained after two function evaluations. They observe that their modified method requires fewer function evaluations than either of the NAG linesearch routines, and that the total for their original method falls between cubic interpolation and quadratic interpolation to the sum of squares. They note occasional inefficiencies in their methods due to extrapolation, but comment that such effects are more pronounced for quadratic interpolation of the sum of squares.

Al-Baali and Fletcher [1986] test similar linesearch methods that use gradients on a set of fifty-five test problems with a number of nonlinear least-squares algorithms described in Al-Baali [1984] (see also Al-Baali and Fletcher [1985]). They conclude that considerable overall savings can be made by interpolating to each of the residuals rather than to the sum of squares. They also obtain favorable results for two different schemes designed to save Jacobian evaluations in the new linesearch.

## 11. Methods for Special Problem Classes

Algorithms have also been formulated to treat some special cases of the nonlinear least-squares problem. For example, there is a vast literature concerning methods specific to nonlinear equations that we shall make no attempt to survey here.

In some nonlinear least-squares problems, the vector  $x$  can be separated into two sets of variables, say

$$x = \begin{pmatrix} y \\ z \end{pmatrix}$$

where it is relatively easy to minimize the sum of squares as a function of  $y$  alone. A fairly common situation of this type is one in which  $y$  is the set of variables that occur linearly in all of the residuals, so that

$$\min_y \left\| f \begin{pmatrix} y \\ z \end{pmatrix} \right\|_2^2$$

is a linear least-squares problem. For example, exponential fitting problems (see Varah [1985]) fall into this category. Methods that deal with separable nonlinear least-squares problems were

introduced by Golub and Pereyra [1973]. Ruhe and Wedin [1980] survey these methods and give some extensions. They describe three basic algorithms, all of which use Gauss-Newton to minimize the sum of squares as a function of  $y$ . The methods differ in the definition of the quadratic model function for minimization with respect to  $z$ . The Jacobian and Hessian of the nonlinear least-squares objective can be partitioned as follows:

$$J = (J_y \ J_z)$$

$$\nabla^2 \left( \frac{1}{2} f^T f \right) \equiv G = \begin{pmatrix} G_{yy} & G_{zy}^T \\ G_{zy} & G_{zz} \end{pmatrix}$$

$$= J^T J + B = \begin{pmatrix} J_y^T J_y & J_y^T J_z \\ J_z^T J_y & J_z^T J_z \end{pmatrix} + \begin{pmatrix} B_{yy} & B_{zy}^T \\ B_{zy} & B_{zz} \end{pmatrix},$$

so that

$$\nabla_z f \begin{pmatrix} y \\ z \end{pmatrix} = J_z^T f,$$

and

$$\nabla_{zz} f \begin{pmatrix} y \\ z \end{pmatrix} = G_{zz} - G_{zy}^T G_{yy}^{-1} G_{zy}$$

$$= (J_z^T J_z + B_{zz}) - (J_z^T J_y + B_{zy})^T (J_y^T J_y + B_{yy})^{-1} (J_z^T J_y + B_{zy}).$$

The approximate Hessians that are considered for the minimization as a function of  $z$  are

$$J_z^T J_z - G_{zy}^T (J_y^T J_y)^{-1} G_{zy}, \quad (11.1)$$

$$J_z^T J_z - J_y^T J_z (J_y^T J_y)^{-1} J_z^T J_y, \quad (11.2)$$

and

$$J_z^T J_z. \quad (11.3)$$

Algorithms based on (11.1) and (11.2) are shown to converge at a faster rate than the conventional Gauss-Newton method, while the asymptotic convergence rate for (11.3) may be much slower. On the other hand, of the three quadratic models, it is least expensive to compute solutions with the approximate Hessian (11.3), and most expensive to compute them from (11.1). Use of (11.2) costs about the same as a conventional Gauss-Newton method. Tests on four sample problems are given to illustrate rates of convergence.

Davidon [1976] introduces a quasi-Newton method for problems in which (i)  $m \gg n$ , (ii) location of the minimum is not very sensitive to weighting of the residuals, and (iii) rapid approach to a minimum is more important than convergence to it. A new estimate of the minimum is computed after each individual residual and its gradient are evaluated, rather than after evaluating the entire block of  $m$  residuals. Davidon gives an analogy to time-dependent

measurements of experimental data, in which quantities calculated from the measurements are updated each time a new observation is made. Starting from an initial quadratic approximation

$$q_0(x) = f(x_0)^T f(x_0) + (x - x_0)^T H_0^{-1} (x - x_0),$$

with  $H_0$  positive-definite, the algorithm that determines the next iterate is equivalent to minimizing a quadratic function of the form

$$q_{k+1}(x) = [\phi_j(x_k) + (x - x_k)^T \nabla \phi_j(x_k)]^2 + \lambda_k q_k(x),$$

where  $\lambda_k$  is in  $(0, 1]$ . It is suggested that the choice of  $\{\lambda_k\}$  should be problem-dependent, and some alternatives are proposed. Davidon tests the method on a set of four problems in which he varies the size of the problem, the initial estimate of the solution, and the sequence  $\{\lambda_k\}$ . He observes that the method tends to oscillate about a minimum rather than converging to it, but that it often reduces the sum of squares more rapidly than other methods.

Further computational experiments with Davidon's method are reported in Cornwell, Kocman, and Prosser [1980]. On a set of fifteen zero-residual problems, they test the method with various fixed values of  $\lambda_k$ . They obtain overflow in most cases for small values, but otherwise find that the efficiency of the method decreases as  $\lambda_k$  is increased. In one case, the method cycled through a sequence of points that was not near-optimal. On the basis of these observations, they implement a new version that attempts to use a fixed, relatively small value of  $\lambda_k$ , restarting from the initial vector with a larger value if it is determined that overflow would otherwise occur. They find that this modified implementation of Davidon's method is competitive with the computer program LMCHOL from Argonne National Laboratory based on Fletcher's [1971] Levenberg-Marquardt algorithm (which has since been superseded by the MINPACK routine LMDER [Moré, Garbow, and Hillstom (1980)]).

## 12. Bibliography

- Al-Baali, M., "Methods for Nonlinear Least Squares", Ph. D. Thesis, Department of Mathematical Sciences, University of Dundee (1984).
- Al-Baali, M., and R. Fletcher, "Variational Methods for Non-Linear Least Squares", *Journal of the Operational Research Society*, Vol. 36 No. 5 (May 1985) 405-421.
- Al-Baali, M., and R. Fletcher, "An Efficient Line Search for Nonlinear Least Squares", *Journal of Optimization Theory and Applications*, Vol. 48 No. 3 (March 1986) 359-377.
- Bard, Y., "Comparison of Gradient Methods for the Solution of Parameter Estimation Problems", *SIAM Journal on Numerical Analysis*, Vol. 7 No. 1 (March 1970) 157-186.

- Bartels, R. H., G. H. Golub, and M. A. Saunders, "Numerical Techniques in Mathematical Programming", in *Nonlinear Programming*, J. B. Rosen, O. L. Mangasarian, and K. Ritter (eds.), Academic Press (1970) 123-176.
- Bartholomew-Biggs, M. C., "The Estimation of the Hessian Matrix in Nonlinear Least Squares Problems with Non-Zero Residuals", *Mathematical Programming* 12 (1977) 67-80.
- Betts, J. T., "Solving the Nonlinear Least Square Problem: Application of a General Method", *Journal of Optimization Theory and Applications* 18 (1976) 469-484.
- Brown, K. M., and J. E. Dennis, "A New Algorithm for Nonlinear Least Squares Curve Fitting", in *Mathematical Software*, J. R. Rice (ed.), Academic Press (1971).
- Cornwell, L. W., M. G. Kocman, and M. F. Prosser, "Computational Experience with Davidon's Least-Square Algorithm", *Journal of Optimization Theory and Applications*, Vol. 31 No. 1 (May 1980) 27-40.
- Davidon, W. C., "Optimally Conditioned Optimization Algorithms without Line Searches", *Mathematical Programming* 9 (1975) 1-30.
- Davidon, W. C., "New Least Square Algorithms", *Journal of Optimization Theory and Applications* 18 (1976) 187-197.
- Dennis, J. E., "Some Computational Techniques for the Nonlinear Least Squares Problem", in *Numerical Solution of Systems of Nonlinear Algebraic Equations*, G. D. Byrne and C. A. Hall (eds.), Academic Press (1973) 157-183.
- Dennis, J. E., "A Brief Survey of Convergence Results for Quasi-Newton Methods", in *Nonlinear Programming*, SIAM-AMS Proceedings Vol. IX, American Mathematical Society (1976) 185-199.
- Dennis, J. E., "III.2 Nonlinear Least Squares and Equations", in *The State of the Art in Numerical Analysis*, D. Jacobs (ed.), Academic Press (1977) 269-312.
- Dennis, J. E., and J. J. Moré, "Quasi-Newton Methods: Motivation and Theory", *SIAM Review*, Vol. 19 No. 1 (January 1977) 46-89.
- Dennis, J. E., and R. E. Welsch, "Techniques for Nonlinear Least Squares and Robust Regression", *Communications in Statistics — Simulation and Computation* B7 (1978) 345-359.
- Dennis, J. E., D. M. Gay, and R. E. Welsch, "An Adaptive Nonlinear Least-Squares Algorithm", *ACM Transactions on Mathematical Software*, Vol. 7 No. 3 (September 1981a) 348-368.

- Dennis, J. E., D. M. Gay, and R. E. Welsch, "ALGORITHM 573 NL2SOL : An Adaptive Nonlinear Least-Squares Algorithm", *ACM Transactions on Mathematical Software*, Vol. 7 No. 3 (September 1981b) 369-383.
- Dennis, J. E., and H. F. Walker, "Convergence Theorems for Least-Change Secant Update Methods", *SIAM Journal on Numerical Analysis*, Vol. 18 No. 6 (December 1981) 949-987.
- Dennis, J. E., D. M. Gay, and R. E. Welsch, "Remark on ALGORITHM 573", *ACM Transactions on Mathematical Software*, Vol. 9 No. 1 (March 1983) 139.
- Dennis, J. E., and R. B. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Prentice-Hall (1983).
- Deuffhard, P., and V. Apostolescu, "A Study of the Gauss-Newton Algorithm for the Solution of Nonlinear Least Squares Problems", in *Special Topics in Applied Mathematics*, North-Holland (1980) 129-150.
- DeVilliers, N., and D. Glasser, "A Continuation Method for Nonlinear Regression", *SIAM Journal on Numerical Analysis*, Vol. 18 No. 6 (December 1981) 1139-1154.
- Eldén, L., "An Algorithm for the Regularization of Ill-Conditioned Least Squares Problems", *BIT* 17 (1977) 134-145.
- Eldén, L., "A Note on the Computation of the Generalized Cross-Validation Function for Ill-Conditioned Least Squares Problems", *BIT* 24 (1984) 467-472.
- Fletcher, R., "A Modified Marquardt Subroutine for Non-Linear Least Squares", Technical Report AERE- R 6799, Atomic Energy Research Establishment, Harwell U. K. (1971).
- Fletcher, R., *Practical Methods of Optimization : Volume I, Unconstrained Optimization*, Wiley (1980).
- Fletcher, R., and C. Xu, "Hybrid Methods for Nonlinear Least Squares", Technical Report NA/92, Numerical Analysis Report, University of Dundee, Department of Mathematical Sciences (December 1985; revised July 1986).
- Fraley, C., "Solution of Nonlinear Least-Squares Problems", Technical Report STAN-CS-87-1165 (Ph. D. Thesis), Department of Computer Science, Stanford University, and Report CLaSSiC-87-04, Center for Large Scale Scientific Computation, Stanford University (July 1987a).
- Fraley, C., "Computational Behavior of Gauss-Newton Methods", Technical Report SOL 87-10, Systems Optimization Laboratory, Department of Operations Research, Stanford University, and Manuscript CLaSSiC-87-21, Center for Large Scale Scientific Computation, Stanford University (August 1987b).



- Fraley, C., "Software Performance on Nonlinear Least-Squares Problems", Technical Report, Systems Optimization Laboratory, Department of Operations Research, Stanford University, and Manuscript, Center for Large Scale Scientific Computation, Stanford University (to appear in 1988).
- Gander, W., "Least Squares with a Quadratic Constraint", *Numerische Mathematik*, Vol. 36 No. 3 (1981) 291-308.
- Gay, D. M., "ALGORITHM 611 : Subroutines for Unconstrained Minimization Using a Model/Trust-Region Approach", *ACM Transactions on Mathematical Software*, Vol. 9 No. 4 (December 1983) 503-524.
- Gill, P. E., and W. Murray, "Newton-type Methods for Unconstrained and Linearly Constrained Optimization", *Mathematical Programming* 7 (1974) 311-350.
- Gill, P. E., and W. Murray, "Nonlinear Least Squares and Nonlinearly Constrained Optimization", in *Numerical Analysis — Proceedings Dundee 1975, Lecture Notes in Mathematics* 506, Springer-Verlag (1976) 134-147.
- Gill, P. E., and W. Murray, "Algorithms for the Solution of the Nonlinear Least-Squares Problem", *SIAM Journal on Numerical Analysis*, Vol. 15 No. 5 (October 1978) 977-992.
- Gill, P. E., W. Murray, and M. H. Wright, *Practical Optimization*, Academic Press (1981).
- Gill, P. E., W. Murray, M. A. Saunders, and M. H. Wright, "Software and its Relationship to Methods", in *Numerical Optimization 1984*, P. T. Boggs, R. H. Byrd, and R. B. Schnabel (eds.), SIAM (1985a), 139-159.
- Gill, P. E., W. Murray, M. A. Saunders, and M. H. Wright, "Model Building and Practical Aspects of Nonlinear Programming", in NATO ASI Series, Vol. F15, *Computational Mathematical Programming*, K. Schittkowski (ed.), Springer-Verlag (1985b) 209-247.
- Gill, P. E., S. J. Hammarling, W. Murray, M. A. Saunders, and M. H. Wright, "User's Guide for LSQOL (Version 1.0): A Fortran Package for Constrained Linear Least-Squares and Convex Quadratic Programming", Technical Report SOL 86-1, Systems Optimization Laboratory, Department of Operations Research, Stanford University (January 1986a).
- Gill, P. E., W. Murray, M. A. Saunders, and M. H. Wright, "Some Theoretical Properties of an Augmented Lagrangian Function", Technical Report SOL 86-6R, Systems Optimization Laboratory, Department of Operations Research, Stanford University (September 1986b).
- Goldfeld, S., R. Quandt, and H. Trotter, "Maximization by Quadratic Hill-Climbing", *Econometrica* 34 (1966) 541-5 1.

- Golub, G. H., and V. Pereyra, "The Differentiation of Pseudo-Inverses and Nonlinear Least-Squares Problems whose Variables Separate", *SIAM Journal on Numerical Analysis* 10 (1973) 413-432.
- Golub, G. H., and C. L. Van Loan, *Matrix Computations*, Johns-Hopkins (1983).
- Gurwitz, C. B., "Sequential Quadratic Programming Methods Based on Approximating a Projected Hessian Matrix", Technical Report # 219, Department of Computer Science, Courant Institute of Mathematical Sciences, New York University (May 1986).
- Häussler, W. M., "A Kantorovich-Type Convergence Analysis for the Gauss-Newton Method", *Numerische Mathematik*, Vol. 48 No. 1 (1986) 119-125.
- Hebden, M. D., "An Algorithm for Minimization using Exact Second Derivatives", Technical Report T. P. 515, Atomic Energy Research Establishment, Harwell U. K. (1973).
- Jones, A., "Spiral — A new Algorithm for Non-linear Parameter Estimation using Least Squares", *Computer Journal*, Vol. 13 No. 3 (August 1970) 301-308.
- Lawson, C. L., and R. J. Hanson, *Solving Least Squares Problems*, Prentice Hall (1974).
- Levenberg, K., "A Method for the Solution of Certain Nonlinear Problems in Least Squares", *Quarterly of Applied Mathematics* 2 (1944) 164-168.
- Lindström, P., and P.-Å. Wedin, "A New Linesearch Algorithm for Nonlinear Least Squares Problems", *Mathematical Programming* 29 (1984) 268-296.
- Marquardt, D. W., "An Algorithm for Least Squares Estimation of Nonlinear Parameters", *Journal of the Institute of Mathematics and its Applications*, Vol. 11 No. 2 (June 1963) 431-441.
- McKeown, J. J., "Specialized versus General-Purpose Algorithms for Minimising Functions that are Sums of Squared Terms", *Mathematical Programming* 9 (1975a) 57-68.
- McKeown, J. J., "On Algorithms for Sums of Squares Problems", Paper 14 in *Towards Global Optimization*, L. C. W. Dixon and G. Szegő (eds.), North-Holland (1975b) 229-257.
- Meyer, R. R., "Theoretical and Computational Aspects of Nonlinear Regression", in *Non-linear Programming*, J. B. Rosen, O. L. Mangasarian, and K. Ritter (eds.), Academic Press (1970) 465-486.
- Moré, J. J., "The Levenberg-Marquardt Algorithm: Implementation and Theory", in *Numerical Analysis — Proceedings Dundee 1977, Lecture Notes in Mathematics*, Vol. 630, Springer-Verlag (1978) 105-116.
- Moré, J. J., B. S. Garbow, and K. E. Hillstom, "User Guide for MINPACK-1", Technical Report ANL-80-74, Argonne National Laboratory (1980).

- Moré, J. J., B. S. Garbow, and K. E. Hillstom, "Testing Unconstrained Optimization Software", *ACM Transactions on Mathematical Software*, Vol. 7 No. 1 (March 1981) 17-41.
- Moré, J. J., "Recent Developments in Algorithms and Software for Trust Region Methods", in *Mathematical Programming—The State of the Art—Bonn 1982*, A. Bachem, M. Grötschel, and B. Korte (eds.), Springer-Verlag (1983) 258-287.
- Morrison, D. D., "Methods for Nonlinear Least Squares Problems and Convergence Proofs, Tracking Programs and Orbital Determinations", *Proceedings of the Jet Propulsion Laboratory Seminar* (1960) 1-9.
- NAG Fortran Library Manual, Vol. 3, Numerical Algorithms Group, Oxford U. K., and Downers Grove, IL U. S. A. (January 1984).
- Nazareth, L., "Some Recent Approaches to Solving Large Residual Nonlinear Least Squares Problems", *SIAM Review*, Vol. 22 No. 1 (January 1980) 1-11.
- Nazareth, L., "An Adaptive Method for Minimizing a Sum of Squares of Nonlinear Functions", *IIASA Report WP-83-99* (1983).
- Nocedal, J., and M. L. Overton, "Projected Hessian Updating Algorithms for Nonlinearly Constrained Optimization", *SIAM Journal on Numerical Analysis*, Vol. 22 No. 5 (October 1985).
- Oren, S. S., "On the Selection of Parameters in Self-Scaling Variable Metric Algorithms", *Mathematical Programming* 7 (1974) 351-367.
- Osborne, M. R., "Some Aspects of Non-linear Least Squares Calculations", in *Numerical Methods for Nonlinear Optimization*, F. A. Lootsma (ed.), Academic Press (1972) 171-189.
- Osborne, M. R., "Nonlinear Least Squares — the Levenberg Algorithm Revisted", *Journal of the Australian Mathematical Society* Vol. 19, Series B (1976) 343-357.
- PORT Mathematical Subroutine Library Manual, A. T. & T. Bell Laboratories, Murray Hill, NJ U. S. A. (May 1984).
- Powell, M. J. D., "A Hybrid Method for Nonlinear Equations", in *Numerical Methods for Nonlinear Algebraic Equations*, P. Rabinowitz (ed.), Gordon and Breach (1970a) 87-114.
- Powell, M. J. D., "A New Algorithm for Unconstrained Optimization", in *Nonlinear Programming*, J. B. Rosen, O. L. Mangasarian, and K. Ritter (eds.) (1970b) 31-65.

- Powell, M. J. D., "Convergence Properties of a Class of Minimization Algorithms", in *Non-linear Programming 2*, O. L. Mangasarian, R. R. Meyer, and S. M. Robinson (eds.), Academic Press (1975) 1-27.
- Powell, M. J. D., "Variable Metric Methods for Constrained Optimization", in *Mathematical Programming—The State of the Art—Bonn 1982*, A. Bachem, M. Grötschel, and B. Korte (eds.), Springer-Verlag (1983) 288-311.
- Ramsin, H., and P.-Å. Wedin, "A Comparison of some Algorithms for the Nonlinear Least Squares Problem", *BIT* 17 (1977) 72-90.
- Reinsch, C. H., "Smoothing by Spline Functions II", *Numerische Mathematik* 16 (1971) 451-454.
- Ruhe, A., "Accelerated Gauss-Newton Algorithms for Nonlinear Least Squares Problems", *BIT* 19 (1979) 356-367.
- Ruhe, A., and P.-Å. Wedin, "Algorithms for Separable Least Squares Problems", *SIAM Review*, Vol. 22 No. 3 (July 1980) 318-337.
- Salane, D. E., "A Continuation Approach for Solving Large Residual Nonlinear Least Squares Problems", *SIAM Journal on Scientific and Statistical Computing*, Vol. 8 No. 4 (July 1987) 655-671.
- Schaback, R., "Convergence Analysis of the General Gauss-Newton Algorithm", *Numerische Mathematik*, Vol. 46 No. 2 (1985) 281-309.
- Shanno, D. F., "Parameter Selection for Modified Newton Methods for Function Minimization", *SIAM Journal on Numerical Analysis*, Vol. 7 No. 3 (September 1970) 366-372.
- Steen, N. M., and G. D. Byrne, "The Problem of Minimizing Nonlinear Functionals", in *Numerical Solution of Systems of Nonlinear Algebraic Equations*, G. D. Byrne and C. A. Hall (eds.), Academic Press (1973) 185-239.
- Stoer, J., "Principles of Sequential Quadratic Programming Methods for Solving Nonlinear Programs", in NATO ASI Series, Vol. F15, *Computational Mathematical Programming*, K. Schittkowski (ed.), Springer-Verlag (1985) 165-207.
- Varah, J. M., "A Practical Examination of Some Numerical Methods for Linear Discrete Ill-Posed Problems", *SIAM Review*, Vol. 21 No. 1 (January 1979) 100-111.
- Varah, J. M., "On Fitting Exponentials by Nonlinear Least Squares", *SIAM Journal on Numerical Analysis*, Vol. 6 No. 1 (January 1986) 30-44.
- Wedin, P.-Å., "On the Gauss-Newton Method for the Non-Linear Least Squares Problem", Technical Report 24, The Swedish Institute of Applied Mathematics (1974).

Wedin, P.-Å., and P. Lindström, "Methods and Software for Nonlinear Least Squares Problems", Report UMINF-133.87, ISSN-0348-0542, University of Umeå, Institute of Information Processing (May 1987).